PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

April 2021

FOREWORD

These proceedings are a written report of the twenty-second Sawtooth Software Conference, held in San Antonio, Texas, April 22-23, 2021. Due to challenges associated with the Covid-19 pandemic, attendance was about one-quarter the usual attendance.

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 optimizing the design and craft of choice/conjoint analysis, surveying on mobile platforms, MaxDiff, market segmentation and classification, and advances in marketing simulations.

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, Eleanor Feit, Joel Huber, David Lyon, and Megan Peitz.

Sawtooth Software

October, 2021

CONTENTS

SUMMARY OF FINDINGS

The twenty-second Sawtooth Software Conference was held in San Antonio, Texas, April 22-23, 2021. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2021 Sawtooth Software Conference Proceedings

Ten Tips for the Effective Presenting of Complex Information (Ray Poynter, NewMR, Potentiate, Nottingham Trent University): Too often good thinking is rendered less accessible by cluttered and ineffective presenting. Ray presented ten ideas to help make presentations more effective, in the context of complex methods, research and thinking. The presentation drew from best practices from sources such as Edward Tufte, Nancy Duarte, Garr Reynolds and Barbara Minto. Ray encouraged us to be human-centric when developing slides; focusing on how people understand and learn, rather than being presenter- or results-centric. Examples included not trying to make the slides become a document, using large enough font sizes, and rounding to two or three significant digits when presenting numeric results. Ray encouraged presenters to focus on the main takeaways, or the key message, rather than having it get lost in the details.

Improving Accessibility for Online Surveys: Looking into Increasing the Online Survey Experience for Individuals with Vision Impairments (Nathan Wiggin, Comengage.US): Fifty-seven million Americans have a disability. Thirty-six million experience challenges with using the internet such as limited dexterity, vision or hearing problems. Nathan and his co-authors pointed out that designing survey research without accessibility in mind can lead to pitfalls including frustration, ostracization, and legal issues. Their presentation recounted the challenges they faced in creating an accessible survey using Sawtooth Software's survey platform (Lighthouse Studio). Screen readers to assist the sight impaired (such as JAWS) can be installed and used to test how understandable and manageable the survey is for people with sight disabilities. The earlier version of Lighthouse Studio led to surveys that were difficult to use for visually impaired respondents. Nathan's team pointed out the problems to Sawtooth Software's developers and made recommendations, many of which were incorporated within a follow-up release of Lighthouse Studio.

Critical User Journey Mad Libs: a new technique for concept development (Jon Gass, Pinterest and Aideen Stronge, Google NBU): As researchers at large organizations, Jon and Aideen described how they often are tasked with prioritizing product requirements and gathering pain points from users. Traditionally this is performed via concept studies in which data are collected on how potential solutions resonate with potential customers. Jon and Aideen introduced a qualitative approach called CUJ Mad Libs for generating the list of customer insights (especially into pain points) that later may be studied more quantitatively using MaxDiff. The approach is inspired by the "Mad Libs" game that involves players filling in their own words to complete phrases. The researcher creates phrases with blanks for the respondent to fill out, such as "[Respondent-provided task name] can be challenging because [Respondent-provided reason]". Other template phrases follow to learn more about the problem and how important it is. Jon and Aideen suggested iterating to create new "Mad Libs" as you discover more information. They suggested the approach could also be adapted for focus groups.

Upgrade Your Brand Tracker Using the Power of Conjoint Analysis (Alexandra Chirilov and James Pitcher, GfK): Alexandra and James reported that most validation studies that compare conjoint preference shares with actual sales only focus on one single point in time. They went one step further and compared conjoint preference shares with real market shares over time. They discussed the ability of a simple brand-price tracker conjoint to capture monthly fluctuations and the long-term trend in brand sales. Their investigation involved three product categories (washing machines, laptops, and TVs) in two countries over 16 monthly waves of data. Because conjoint analysis involves showing products at realistic prices, they found that conjoint preference was superior to traditional stated brand preference in tracking with actual purchase shares. Calibrating conjoint predictions by overall distribution effects improved conjoint's predictions of actual market shares, but adjustments for brand awareness in the conjoint market simulator did not. They also found that conjoint analysis trackers were more stable across waves and also more correlated with long-term trends in purchase shares compared to stated brand preference. However, neither conjoint analysis nor stated brand preference were able to predict short-term fluctuations in market share. Alexandra and James pointed out there are other factors that affect market shares such as in-store promotions, product placement, and sales staff effectiveness. Such effects cannot be captured well by either stated brand preference or conjoint analysis and this may have explained why short-term fluctuations in market share were not predictable.

Respondent Quality: Identifying Bad Respondents from MaxDiff Response Patterns (Jane Tang, Mona Foss, and Rosanna Mau, Bootstrap Analytics): Jane and her co-authors extended previous research regarding how to reliably identify bad respondents in MaxDiff surveys. They specifically examined three different approaches to identifying random patterns of responses: HB's RLH fit score, and a logistic regression approach involving both HB's RLH score along with either Latent Class (LC) and Scaled-Adjusted Latent Class (SALC). With the HB RLH fit criterion, random respondents are generated and a fit cutoff is selected based on the distribution of RLH scores for these random respondents. The two Latent Class approaches involved finding a group (or groups) of respondents who seemed to reflect random responses. Jane and her co-authors looked at how removing suspect random responders using the three methods would improve the differentiation in the MaxDiff scores as well as improving the differentiation in concept test rating results. They concluded that the simple HB RLH approach and the more complex RLH + SALC performed about equally well for identifying bad respondents and improving the signal-to-noise ratio in the data. Because the HB RLH approach is much simpler, they recommended this approach for practitioners.

Uncommon Choices: Novel Applications of Conjoint Analysis in Practice (Chris Chapman, Google): Chris reviewed four novel applications of Conjoint Analysis for product design from various presentations he's delivered over the past 12 years at Sawtooth Software Conferences. He highlighted how each method was useful, describing its key points for success. The four applications were as follows: 1) comparing CBC and ACBC for real respondent preferences, where ACBC was found to slightly improve predictions of real choices, 2) using Game Theory with conjoint analysis, to decide whether to add a product enhancement or not while considering the possible reactions of a key competitor, 3) using genetic algorithms with market simulators to figure out which products to include in a streamlined product line, and 4) using "Profile CBC" for psychographic segmentation, where respondents pick which conjoint profile most resembles them in terms of personality characteristics. When comparing CBC and ACBC results, Chris found that they were very similar, but that ACBC seemed to have a bit better precision. He recommended ACBC for smaller sample sizes and when you can take some extra time with respondents in the interview. As for "Profile CBC," Chris recommended its use for developing personas (characteristic profiles of groups of buyers) via latent class that validate with external measures and directly follow from the CBC data. In conclusion, Chris stressed that conjoint analysis works very well and that the cost of collecting the data and performing the research outweighs the costs of making business mistakes.

Comparing Predicted Automotive Purchase with Passive Geolocation Covariates to Actual Ownership Records (Marc Dotson, BYU and Edward Paul Johnson, Harris Poll): Paul and Marc investigated the use of passive geolocation data regarding visits to automotive service locations as covariates in HB modeling for a standard CBC model to predict car purchase. Passive geolocation data eliminates recall error, since it directly observes whether respondents visited different car dealerships. However, there are challenges with geolocation data involving geographic measurement error and issues with smartphone availability/permissions. Paul and Marc examined the usefulness of different covariates in their HB modeling for their CBC data. Those covariates included geolocation data for dealership visits, demographics, and stated preferences for car type, price, and brand. They found that holdout hit rate predictions for CBC choices using covariates improved to 0.50 from 0.36 when leveraging all three types of covariates, though most of gain in hit rate could be accounted for by stated preference data and demographics. They had planned to be able to do more with actual purchase information via follow-up with their original respondents, but ran into various troubles largely due to the Covid pandemic in assembling the validation data.

*** Enhance Conjoint with a Behavioral Framework** (Peter Kurz and Stefan Binner, bms Marketing Research + Strategy): Finding useful covariates (variables outside the conjoint task) that can boost the predictive validity of conjoint analysis via HB estimation is a topic that Peter and Stefan have investigated and reported on multiple times at past Sawtooth Software conferences. In this presentation, Peter and Stefan found better success than past investigations by using a simple set of nine questions (binary semantic differentials) that can be added to the survey just prior to the CBC questions. The nine questions are based on principles of behavioral economics and include such pairs as: “I think brands differ a lot” vs. “I think brands are more of less the same” (respondents pick the statement they most agree with). Peter and Stefan proposed that these simple pairs statements help respondents remember their prior shopping situations and prime them to do a more realistic job in answering CBC questions. The showed that hit-rates and out-of-sample predictions could be significantly improved using this framework. They presented results for nine different CBC studies, demonstrating good improvement in both in-sample and out-of-sample hit rates when leveraging the nine covariates. What was perhaps even more intriguing is that just the mere presence of the covariates seemed to improve respondents’ performance on the CBC tasks. Even without using the covariates in HB estimation, the act of completing the nine semantic differential pairs improved the predictability of the respondent’s utilities for out-of-sample holdouts. In addition to the usefulness of covariates for improving predictive accuracy of the conjoint results, Peter and Steven recommended their value in developing useful consumer segmentations.

*Best Paper Award, based on audience evaluation and steering committee recommendation.

Enhancement of Van Westendorp Price Model via Newer Statistical Approaches (Ming Shan, KYNETEC): In this presentation, Ming described some of the methodological shortcomings of the frequently used Van Westendorp (PSM) price model (the “4 pricing questions” approach). He also highlighted the Newton-Miller-Smith (NMS) extension that asks for purchase intent on the two middle price points. Ming forwarded the idea of applying survival analysis to PSM, where price was used instead of the traditional time variable. A key change that Ming suggested for his models was that probability of choice (via the NMS extension) should be considered to be 1.0 at the respondent's stated lowest price rather than 0.0. Ming found that the optimal price suggested by PSM and the NMS extension were higher and less stable than those developed from the new models. Furthermore, the new models could be used to develop confidence intervals via the survival model framework. He described how he is developing an R package for sharing that implements the models described in his paper.

Estimating Willingness to Pay (WTP) Given Competition in Conjoint Analysis (Bryan Orme, Sawtooth Software): Since the inception of conjoint analysis, clients have wanted to use the data to figure out how much consumers are willing to pay for enhanced features. Bryan described how the two most common approaches to WTP (the algebraic approach and the two-product 50/50 simulation approach) ignore competition and tend to overstate WTP. He reviewed an intuitive approach to WTP from decades ago involving simulating the client's product versus a rich set of competitors and the None alternative. The client's product is enhanced and the price that drives its share back to the original base case share is taken as the WTP. Bryan showed that the WTP results using the competitive set approach tended to be lower than other traditional approaches over nine CBC datasets. He introduced a new approach called Sampling of Scenarios (SOS) when the base case client and competitor products isn't well known. SOS involves randomly picking attribute levels for the client's product as well as competitors. SOS can involve constraints on the products such as accounting for the client's brand (while competitors cannot take on that brand name) or accounting for the client having a patent on a product feature. Bryan also described how confidence intervals on the WTP estimates could be developed using bootstrap sampling.

Filter CBC: A New Approach to Mimic the Online Shopping Experience (Marco Hoogerbrugge, Menno de Jong, Kevin Lattery, and Kees van der Wagt, SKIM): Marco and his co-authors showed a new approach to conjoint analysis that uses dozens of concepts per task and allows respondents to filter and sort the concepts like they would do in online shopping. Filters allow respondents to narrow the product selection to acceptable ranges of features and price. Sorting allows respondents to prioritize the acceptable options on more important features. In the end, the respondent makes a single choice from dozens of concepts in each task; but the filtering selections could also be used in modeling the data. Marco described an experiment they conducted in two waves: first a conjoint survey for mobile phone plans (involving both filter CBC and standard CBC tasks), then later they asked the same respondents what phone plan they had purchased. They compared conjoint predictions with reality. They found that the filter-CBC tasks provided slightly better predictions of later actual purchase behavior than the standard CBC tasks. They also reported their results by different ways of including the filter selections in the CBC modeling. Ignoring the filter information led to very good models, but a fancier conjunctive screening model using the filters information could lead to potentially better results.

Using MaxDiff instead of CBC? Pros, Cons, and Recommendations (Megan Peitz and Abby Lerner, Numerious): With the introduction of Anchored MaxDiff from Jordan Louviere (Indirect Anchoring) and Kevin Lattery (Direct Anchoring), we're given a new set of capabilities and a proxy for the "none" alternative that we previously only had in a Choice-Based Conjoint. Megan and Abby examined whether we could take smaller CBC designs (e.g., four attributes each with three levels) and turn them into Anchored MaxDiffs (by taking the 3^4 full-factorial combinations as a list of 81 items into a sparse MaxDiff). Or, sometimes our conjoint designs have so many alternative-specific relationships or prohibitions that the full-factorial of concept combinations could lead to 120 or fewer total concepts that could be treated as items in a sparse MaxDiff. Benefits of using MaxDiff in this way could include automatically capturing all higher-order interaction effects, easily running TURF analysis, and creating clustering solutions based on reactions to the entire product preference, not just individual feature preference. Megan and Abby fielded a five-cell experiment that tested different ways of asking CBC questions and MaxDiff questions with the None alternative to compare the results. They found significant differences in the None prediction from CBC compared to direct anchored MaxDiff questions. Although using a sparse MaxDiff covering the full-factorial for small conjoint designs seems like a viable approach, they concluded that it did not perform better than CBC for predicting in-sample and out-of-sample holdout choices.

Concept Screening and Evaluation Using Survey Based Artificial Markets (John Pemberton, Alfred Johnson, Emily Gasparro, Alyssa MacDonald, and Lauren Piskorski, Butler University and Indiana University): Faced with the challenges presented by established concept testing and screening techniques, John and his co-authors described an alternative approach that utilizes an artificial market in which respondents evaluated 16 wholistic product concepts on a 0-100 scale. The algorithm they used rewards respondents based on the quality of their judgments, which creates the feeling of a game or competition, and the winner gets some cash to buy articles in the university bookstore. The artificial market algorithm tested different prices for the concepts to find the balance where there were about as many people who indicated a “buy” vs. a “sell”. The research team fielded a standard MaxDiff survey for the 16 product concepts as well as implementing the prediction market experience. Due to challenges on campus with Covid, the number of participants was not as many as initially hoped for. Even so, the correlation between MaxDiff and the prediction market approach was above 0.7. In addition to sampling error being a possible explanation of some differences in rankings between MaxDiff surveys and prediction markets, another possible difference was that respondents in the prediction market were trying to evaluate the concepts based on what they thought the market receptivity would be, rather than their own personal preferences.

ACBC vs. Partial Profile CBC: A Market-Based Comparison (Fisher Liu, Diagaid, Shumin Wang, Yi You, Vivo Mobile Communication CO, and Dapeng Cui, Diagaid): Partial-profile CBC is an alternative to ACBC when there are many attributes in conjoint studies. In this presentation, Fisher and co-authors compared partial-profile CBC to ACBC for a smartphone choice study involving 14 attributes. The utility estimates and sensitivity analyses from both methods were similar, but partial-profile seemed to neutralize the differences between attributes relative to ACBC. A benefit for partial profile is that it took less time for respondents to complete than ACBC (two-thirds as long). However, ACBC was able to detect and better measure an important interaction between RAM and ROM attributes, whereas partial-profile was not able to detect the interaction. That particular tradeoff (between RAM and ROM) was critical to the client’s design decision for a new product launch, so a second-wave survey was fielded which isolated just the RAM and ROM tradeoff. That follow-up study demonstrated that the first-wave ACBC model better predicted respondents’ tradeoffs between RAM & ROM than the partial-profile prediction. With that second-wave survey validation in hand, the client decided to rely on ACBC’s prediction regarding the optimal combination of RAM and ROM to offer. Later sales data showed that ACBC’s predictions very closely predicted the actual sales result. The client launched two products with different RAM & ROM specifications. Relative to the key competitor, the ACBC predictions for the three products were 33%, 21%, and 46%; and the real relative market shares after launch ended up being 31%, 20%, and 49%.

Replication of Known Segment Structure and Membership: A Data-Driven Comparison of Robust Partitioning Methods for Metric Data (Keith Chrzan, Sawtooth Software and Joseph White, InMoment): When segmenting the market using only metric input variables, several algorithms could be used. Because "truth" isn't typically known in segmentation exercises, we don't know which of them works best. Relying on artificially constructed data sets with known segmentation solutions, Keith and Joseph tested a handful of segmentation algorithms to see which fared best in terms of identifying the correct number of segments and in correctly assigning respondents to segments. The test data sets were experimentally created to reflect a number of varying data conditions (number of segments, relative segment sizes, degree of segment separation, number of dimensions and number of variables per dimension). Keith and Joseph tested five methods: PAM, k-means (Sawtooth Software's CCA), cluster ensembles (Sawtooth Software's CCEA), k-means with 1000 starting points (in R), and Latent Class clustering (in R). Additionally, Keith and Joseph used various approaches to measuring quality of solutions to see if the correct number of clusters could be identified reliably (Hopkins statistic, NbClust ensemble of cluster measurements, Sawtooth Software's reproducibility, and Silhouette statistics). They found that the K-means package with 1000 starting points in R as well as Sawtooth Software's CCEA package did the best in recovering true membership in segments (assuming the analyst knows the correct number of segments). However, none of the fit statistics were able to point very successfully to the true number of segments for the artificial data sets they constructed. They concluded that researchers should be skeptical of statistics meant to point to the correct number of segments and be willing to leverage client expert knowledge regarding how many segments make sense for their business problems.

Bi-Cluster Identification & Profiling (Ewa Nowakowska, EY and Joe Retzer, ACT Market Research Solutions): Traditional data clustering algorithms are challenged both by conceptual as well as practical issues. In this presentation, Ewa and Joe presented an approach, bi-clustering, which addresses the presence of dyadic relationships in clustering data. The demonstrated how this method can profile the resultant "cluster cells" graphically. With bi-clustering, subsets of respondents cluster on different subsets of features. Respondents can be in more than one cluster, but one can constrain the algorithm so that each respondent is assigned to only one cluster. Similarly, one can constrain bi-clustering so that features only group with one cluster. Bi-clustering can use multiple algorithms and different profiling data (binary, ordinal, metric). Ewa and Joe pointed out that bi-cluster solutions in which respondents uniquely belong segments are easier for clients to understand; though clients can easily handle the notion that features are not mutually exclusive to clusters. They demonstrated how bi-cluster analysis could be applied to a data set of new auto buyers, where respondents have been asked to indicate (binary response) regarding which of 45 reasons for buying are important to them. To help determine the appropriate number of clusters, Ewa and Joe suggested a bootstrapped approach that replicates the solution many times and finds the agreement across replicates using the Adjusted Rand Index. As caveats, they mentioned that using the BCBimax algorithm leads to long run times and it doesn't ensure that all observations are classified into a cluster.

Conjoint Analysis for Difficult Choices (Joel Huber, Duke University and Martin Meissner, Zeppelin University): Joel and Martin expressed based on their experience that conjoint analysis sometimes faces problems measuring and predicting people's choices for certain challenging topics. Problems occur when the attributes themselves are hard to understand, when the decision is novel, if the tradeoffs can involve unacceptable levels, and if the decision is irreversible. Some of these problems can be addressed by recruiting the right people, introducing each attribute level and relating it to their lives, and using simple labels or icons to describe them. With difficult decisions, especially involving such topics as cancer treatment or hand replacement surgery, Joel and Martin expressed that asking pairs of choices is better than triples or quads. They described a conjoint study involving hand replacement (after the loss of a hand), where the results of conjoint analysis pretest involving 100 respondents didn't work very well to predict people's final choices between a prosthetic hand and hand replacement surgery. In a follow-up study, they conducted a conjoint study involving only options for a prosthetic hand. Then, they provided more details about the hand transplant process and asked respondents to identify statements regarding concerns and important features involving that decision. Respondents were asked to then make a choice about transplant or not; then depending on their choice a new set of new statements (represented to be from physicians) were shown to them either for or against transplants. The respondents then repeated their choice, given that new information purportedly from physicians. Joel and Martin found that this process helped respondents better understand the difficult decision and provide better insights regarding their probable choices.

Concordance between Treatment Choice and Preference for Localized Prostate Cancer (Ravishankar Jayadevappa, Sumedha Chhatre, Joseph J. Gallo, Marsha Wittink, Knashawn H. Morales, David I. Lee, Thomas J. Guzzo, Neha Vapiwala, Yu-Ning Wong, Diane K. Newman, Keith van Arsdalen, S. Bruce Malkowicz, and Alan J. Wein): Treatment choice for localized prostate cancer is preference sensitive given that several medically viable and effective treatment options are available, each with specific risks and benefits. Ravishankar and co-authors reported on the effectiveness of a PreProCare preference assessment instrument (using adaptive conjoint analysis) to assess the association between value markers (utility levels) and treatment choice in localized prostate cancer patients. They observed that preference assessment intervention helped prostate cancer patients reveal their preferences, leading to better alignment with treatment decision. When patients completed the PreProCare preference assessment questionnaire, they received real-time recommendations regarding their most important concerns (attributes) to discuss with their physicians. This facilitated enhanced discussions between patient and physicians regarding the most appropriate treatment for their stage of prostate cancer. Patients were randomly selected to receive the PreProCare instrument (n=371) or not (the control group, n=372). After 24 months, the PreProCare group showed greater satisfaction with care and satisfaction with their decision than the control group. Also, among low-risk patients, a higher proportion of the PreProCare instrument group were on active surveillance rather than more aggressive treatment options. For the PreProCare group, treatment choices were more consistent with the estimated risk level of prostate cancer diagnosis.

Price Fixing: Can you construct an experimental design to make price sensitive consumers appear price seeking? (Jake Lee, Red Analytics): The impact of experimental design on the inferential results of a choice experiment is often overlooked. Jake expressed that designs are the most complicated and least sexy part of an experiment. When software makes it easy to generate the experimental plan, researchers sometimes ignore deficiencies that can lead to poor outcomes. Jake demonstrated what can go wrong if you use a poor design (e.g., too many prohibitions) and fail to pay attention to design diagnostics that would suggest a problem. He also showed how overly constrained designs can lead to problematic repetitions of pairs of attribute levels within choice tasks (price-dominated brand pairs). The particular data set Jake used for illustration led to a great many respondents with positive coefficients for price. He stressed that the number of things to consider in choice experiments (e.g., level balance, level overlap, orthogonality, utility balance, appropriate complexity of the attribute list, and how many times each level appears within each respondent's tasks) means that we should pay greater attention to the process rather than just trust that a design algorithm has done an adequate job. Since manual inspection of designs to root out problems is tedious, Jake recommended automated procedures for detecting them.

Should Shapley-Adjusted Regression Coefficients Be Used in Making Predictions? (Jack Horne, Jack Horne Consulting): Shapley regression is a useful tool for understanding relative importance of predictors. For example, a company might be interested to know what attributes about a customer's experience with the company lead to outcomes such as satisfaction or loyalty. With such information, not only would companies understand the importance of factors for driving satisfaction, but they'd be able to predict gains in satisfaction due to making improvements in specific aspects of their business. Jack pointed out that due to collinearity in the independent variables, standard regression weights can be unreliable—even displaying reversals (negative coefficients) when we know from pairwise correlations that improvements on such variables are associated with positive outcomes in the dependent variable. Methods to deal with the problems of collinearity among the independent variables include applying principal components analysis, but Jack suggested an approach called Shapley regression. Shapley values are based on how much each predictive variable adds to the dependent variable outcome when involved in coalitions (all possible sub-groupings) of predictive variables. Jack stated that although Shapley values give a better read on the true contribution of each variable to the dependent variable outcome, they are not regression coefficients and should not be used in prediction. He reported that it is possible to derive predictive regression coefficients by optimizing on the Shapley values (and suggested a routine in R and in Python to do so). But, he demonstrated for a particular data set that the fit of the Shapley adjusted coefficients is less than the traditional regression analysis and the coefficients were flatter. Jack recommended that if the purpose of the model is prediction, using the standard regression coefficients should be preferred. But, if the goal is to interpret individual-level coefficients, then the Shapley value coefficients would be preferred.

An Integrative Model for Complex Conjoint Analysis (Yichun Miriam Liu, Ohio State University, Jeff Brazell, University of Utah, and Greg Allenby, Ohio State University): Miriam and her co-authors focused on how to improve conjoint analysis when dealing with large numbers of attributes: situations in which a complex product is made up of simpler, component products. While many academics discount the widespread need for such models, they showed a result from a Sawtooth Software survey of its customers in which 36% of conjoint analysis studies were reported to have 10 or more attributes and 12% involved 15 or more attributes. A challenge in modeling demand for complex products lies in simultaneously studying features that affect component preference and its resulting effect on marketplace demand. They argued that there are differences in modeling preference for the whole product versus modeling choices that buyers make regarding the components of an offering. They proposed an integrative model for studying complex products utilizing multiple conjoint exercises within a single structure. A key aspect of their model was including price as an attribute in the overall product choice as well as an aspect of the choice of components within the product. Miriam and her co-authors noted that they are not the first to suggest that multiple conjoint studies could be bridged via common attributes, but that their model accounts for differences in scale (response error) between the overall and sub-component conjoint analysis designs. They illustrated their model using a conjoint study of demand for option packages when purchasing an automobile.

TEN TIPS FOR THE EFFECTIVE PRESENTING OF COMPLEX INFORMATION

RAY POYNTER
NEWMR, POTENTIATE, NOTTINGHAM TRENT UNIVERSITY

ABSTRACT

It does not matter how good your analysis is, if you do not communicate the results effectively your impact will be less than it could be. Indeed, a failure to communicate adequately could undermine your work entirely. In this paper I set out ten ways that you can increase the effectiveness your presentations when sharing complex information.

INSPIRATION

The inspiration for this paper is drawn from watching a series of presentations from the late Hans Rosling, who was a truly great communicator. If you are not familiar with his work, I suggest you Google "Hans Rosling TED" and watch a selection of his TED presentation recordings. One great example to start with is "The best stats you've ever seen," presented at TED 2006.

THE TEN TIPS

In this paper I set out ten tips that will help you make your communications more effective. These tips do not focus on engagement, narrative, or visualization—they focus on simple steps that you can take to improve the chance that your message will be understood. The tips are:

1. Make it readable—to everybody
2. Context is king
3. A presentation is not a slideument
4. Sort the data by something meaningful
5. Less is more
6. Use comparisons
7. Charts should not NEED numbers
8. Convey the message not precision
9. Don't bury the story
10. Dale Carnegie in modern form

In the sections below I outline each of these ten tips, adding examples where appropriate.

1. MAKE IT READABLE—TO EVERYBODY

The first tip is to simply make sure that people can see what you are presenting. The first consideration in making your presentation readable is to choose a large enough font size. The marketing guru Guy Kawasaki has offered an algorithm for determining the font size to use with an audience, he suggests "*find out the age of the oldest person in your audience and divide it by two*." If your audience has 50-year-olds in its number, use a 25-point font. In a more general sense, go to the back of the room and view your presentation; is everything that should be read truly visible? If you are presenting virtually (via Zoom or Teams), check what works for the sorts of screens that your audience is using—it may not be the same as on your 28-inch widescreen monitor.

Another readability consideration relates to color blindness. About 8% of men have some color deficiency (and a smaller proportion of women). If there are ten men in your audience, there is a good chance at least one of them will struggle to see the difference between red and green—and it is unlikely they will tell you this in advance. To make your presentations more readable, to more people, use color and shading to highlight information. You can also choose a color pallet that is flagged as being color-blind safe. Color blind pallets use tones/intensity to help differentiate the colors.

If you are making your presentation available to others to read later, do not forget to put Alt-tags on your images. People with visual challenges may be accessing your presentation via a screen-reader (a device with turns the text of your presentation into sound), these people need your charts and images to utilize Alt-tags. Note, you can use Alt-tags in PowerPoint, in Keynote, in Google Slides, as well as in word processing programs such as Word.

2. Context Is King

A good presentation is a function of the audience, the message, and the presenter—there is no single best way to present, it depends on the context of these three elements. In terms of the audience, what do they already know, what are they expecting to hear, how long are they available for, what styles of presenting work well with them, and what sorts of "proof" do they respond to? In terms of "proof," some audiences want logic, some want case examples, some want quantitative evidence, some want qualitative insight. Providing somebody with the wrong type of evidence is the same as not providing evidence.

In terms of the message, one key consideration is whether it is "good news" or "bad news." If you are presenting good news to an audience (for example, that their proposed project is likely to succeed), you are likely to find this an easy presentation. However, if you are presenting bad news (for example, telling them that their much-loved project is likely to fail) then the audience is likely to challenge your evidence and conclusions.

The final ingredient is you. Different presenters have different strengths and weaknesses. If you are a recognized expert in a field, you may need to show fewer workings and justifications. However, if you are relatively unknown you may need to provide more support for your observations and conclusions. If you are confident in your knowledge of the material, you might have very little text on the page, if you are less confident then you might want more supporting material.

3. A Presentation Is Not a Slideument

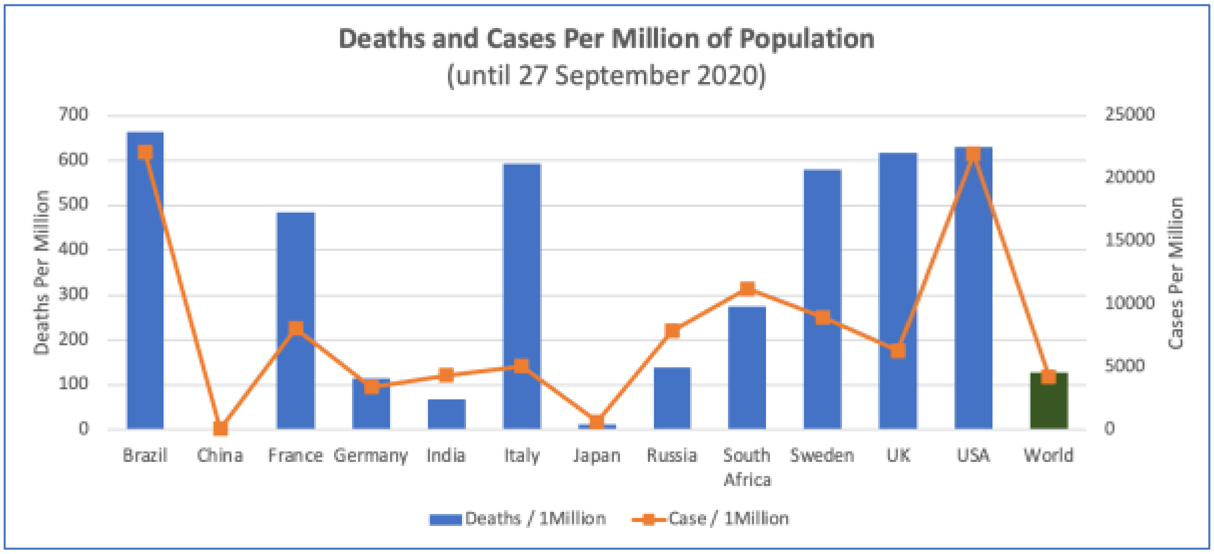
The presentation guru Garr Reynolds (and author of the book and website PresentationZen) uses the term slideument to describe the process of trying to create something which is both the material shown during a presentation and the material that is left behind/provided as a reference document. In almost all cases, a good presentation is not a good reference document, and in almost all cases a good reference document is not a good presentation.

One relatively straightforward option for creating a reference document from a presentation is to fully utilize the Notes facility on packages such as PowerPoint—and then ensuring that the Notes view is the format used for the "leave behind."

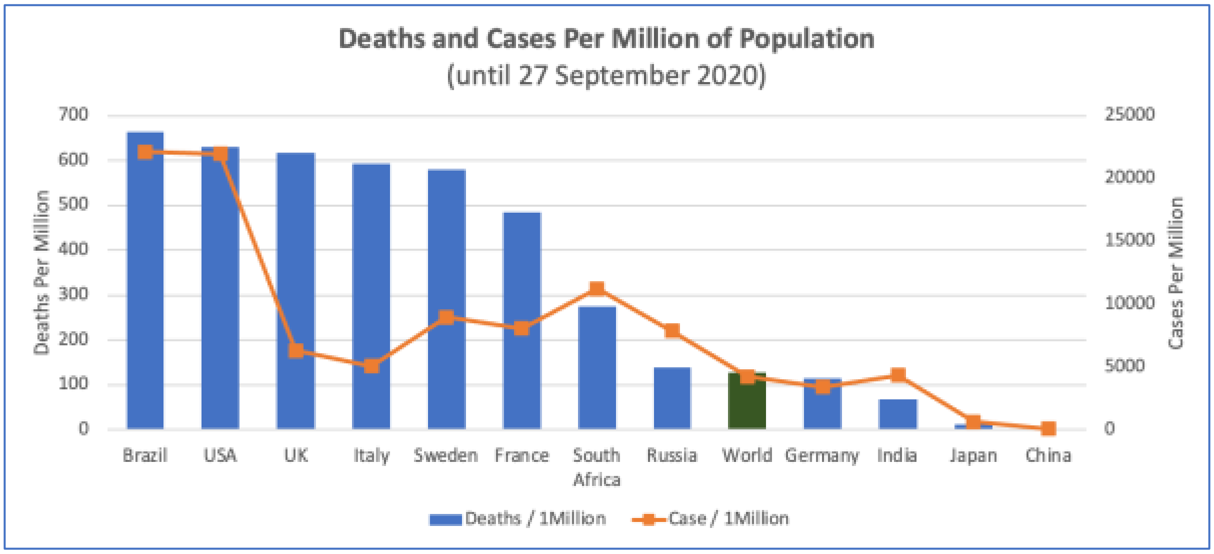
4. Sort the Data by Something Meaningful

Too many presentations show the data in an order that is not designed to help the audience grasp the message being communicated. Tables and charts are too often shown in the order they were listed in a questionnaire or in the order they appeared in some online source, or the order they appeared in the output from an analytics package. For example, the results of a cluster analysis might display four clusters, but in the arbitrary order the software created them in, rather than in a sequence that helps communicate the information.

The two charts below illustrate this point with data downloaded from the Worldometer site, relating to the Coronavirus pandemic. The first shows the data alphabetically, which was the order the website supplied the data in.



The problem with this chart is that the "message" is not highlighted. To highlight the message, we can simply re-order the countries in terms of the number of deaths per million. This sorted view is shown in the second version of the chart, shown below.



By sorting the data by deaths per million, we can see that there is a substantial mismatch between the number of cases identified per million and the deaths per million in some countries. The death figures are generally deemed to be broadly accurate; the case figures are a function of how many people were tested. The chart shows that in Germany, Brazil and the USA the ratio of cases and deaths were broadly in line with the ratio for the whole world. By contrast, the UK, Italy, Sweden and France appear to be testing fewer people, creating a different impression of the link between cases and deaths.

5. Less Is More

When people asked the Renaissance sculptor Michelangelo how he created the famous statue of David, he is said to have answered "*It is easy, you chip away anything that is not David.*" This is a message that applies to presentations and reports, chip away anything that is not the message.

The first element of "less is more" is to look at your material and check whether there is anything distracting from or irrelevant to your message. Can you reduce the number of words, numbers, columns, images and leave the message 99% intact? If so, remove them. In a "leave behind" you might want your logo on every page, but in a live presentation it is probably a distraction.

One of the ways to create less is to use fewer significant digits. As Stanislas Dehaene shows in his book, *The Number Sense*, humans are not very good at processing numbers with multiple digits. We can make our messages clearer by using fewer digits, for example using two or three significant digits. The table below shows several examples of using two and three significant places.

Raw	2 SD	3 SD
Cost = $36,723	$37K	$36.7K
R-squared 45.67%	46%	45.7%
Height 1751mm	1.8m	175cm
Time 2 hour 32 minutes 10 seconds	2.5 hours	152 minutes

Moving on from significant digits, it is worth considering decimal places and percentages—especially in the context of data based on samples (where the base is often 1000 people or less). Consider an example where study has looked at the preferences of two cells of 1000 people, one cell evaluating A and one cell evaluating B. The research finds that:

- A is liked by 50.47% of one cell
- B is liked by 49.68% of the other cell.

If we report the data to two decimal places, it looks to most readers as if we are saying that A is bigger than B. But, we know that the sampling error of these two estimates is likely to be +/- 3%. What we mean is that A and B are both about 50% (plus or minus 3%).

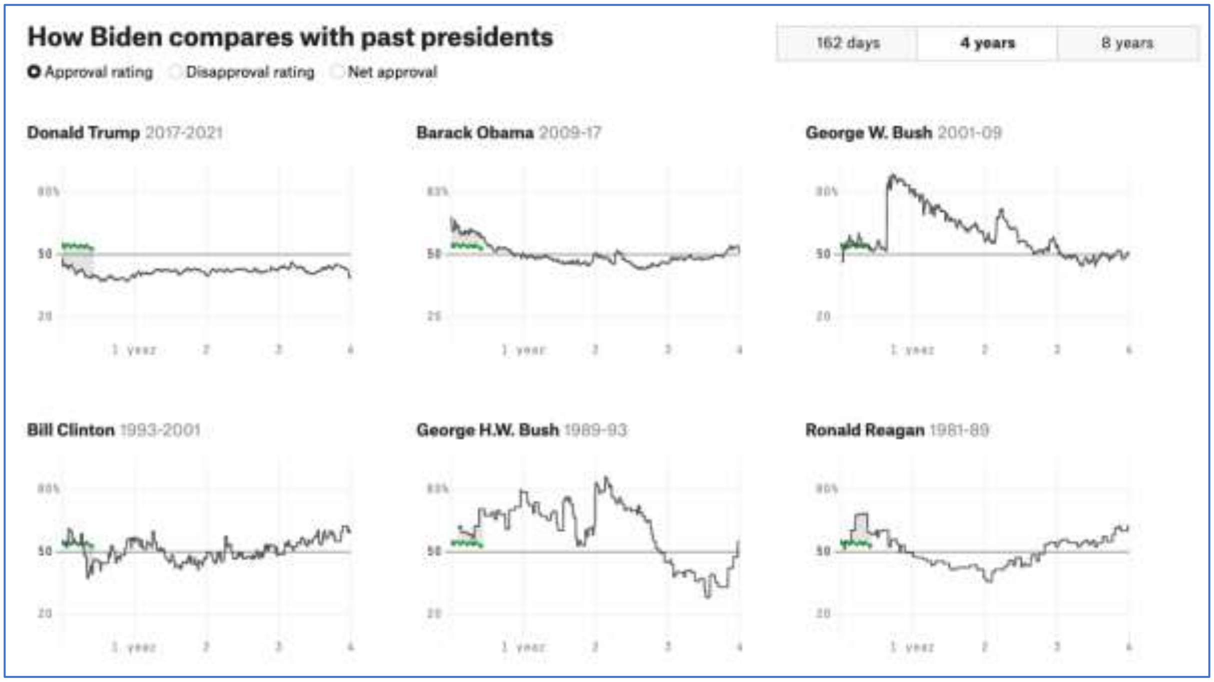
Decimal places make numbers harder to interpret, in this case sticking to two significant digits will help the casual reader come to the right understanding of the message.

There are times when small differences matter, and those differences might be expressed as percentages. For example, consider the deaths per million data in terms of the COVID pandemic. Up to 2 April the number of deaths in the USA per million of population was 1715, which is a rate of 0.17%. In New Jersey there had been 2771 deaths per million, a rate of 0.28%. The difference between the two numbers is 0.11 percentage points. These numbers are important and the difference is important, and they are based on a census, not a sample. To make the numbers clearer, we can use the system that finance uses, namely basis points. There are 100 basis points in a 1 percentage point difference. This means we can express the difference between New Jersey and the USA as 11 basis points—which focuses on the difference, not on things like decimal places.

6. Use Comparisons

Most new numbers need a comparator if the audience is going to appreciate their meaning. For example, I recently took part in a 108-mile race, raising money for charity. My son posted about my race as part of the fundraising and highlighted that this was the same as running four marathons in one event and then running four more miles. For non-ultra-runners this gave more context.

One way that we often leverage comparisons is in terms of benchmarks. For example, if we are testing a new TV commercial, we might report that it was in the top decile of all ads tested. Nate Silver's FiveThirtyEight.com site often features great visualizations of data. The charts below show the approval ratings for President Biden and compares them with other recent presidents—allowing President Biden's figures to be assessed in a relevant context.

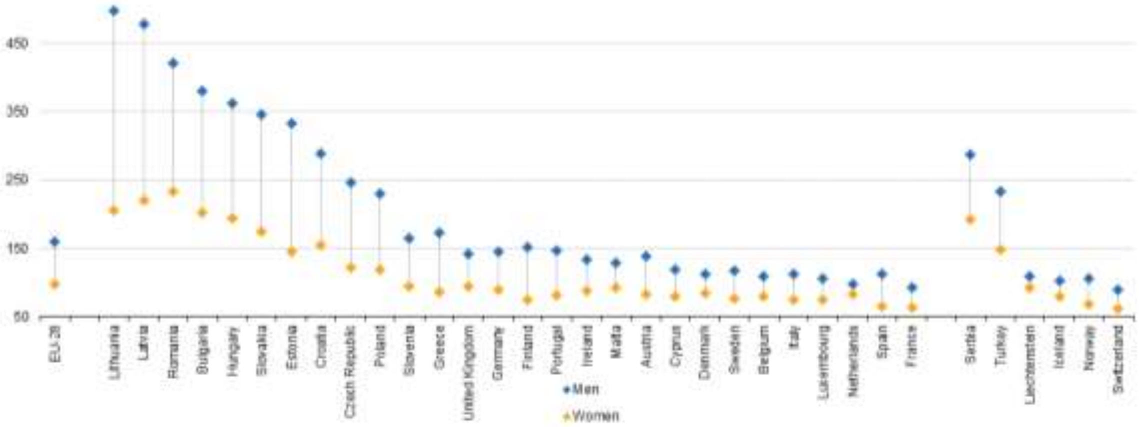


With comparisons, care should be taken to ensure that they actually help envision the meaning of the numbers. For example, to be told that something is the same size as four school buses is quite meaningful, but to be told it is as large as 2000 school buses is less helpful (most of us would struggle to tell the difference between 1500, 2000, and 2500 buses).

7. Charts Should Not NEED Numbers

Note, charts can have numbers, especially if the client asks for them, but my point is that charts should not NEED numbers. If a chart needs lots of numbers, then usually it means the chart is not clear enough, or it means that the user needs a table, not a chart.

At the start of the paper, I referred to Hans Rosling and his amazing data presentations. If you watch one or several of his TED recordings you will note that he displays very few numbers—he ensures the meaning is clear without the numbers. The chart below is from a publicly available report about Amenable Deaths in Europe (amenable deaths are deaths that could have been avoided if the right steps had been taken).



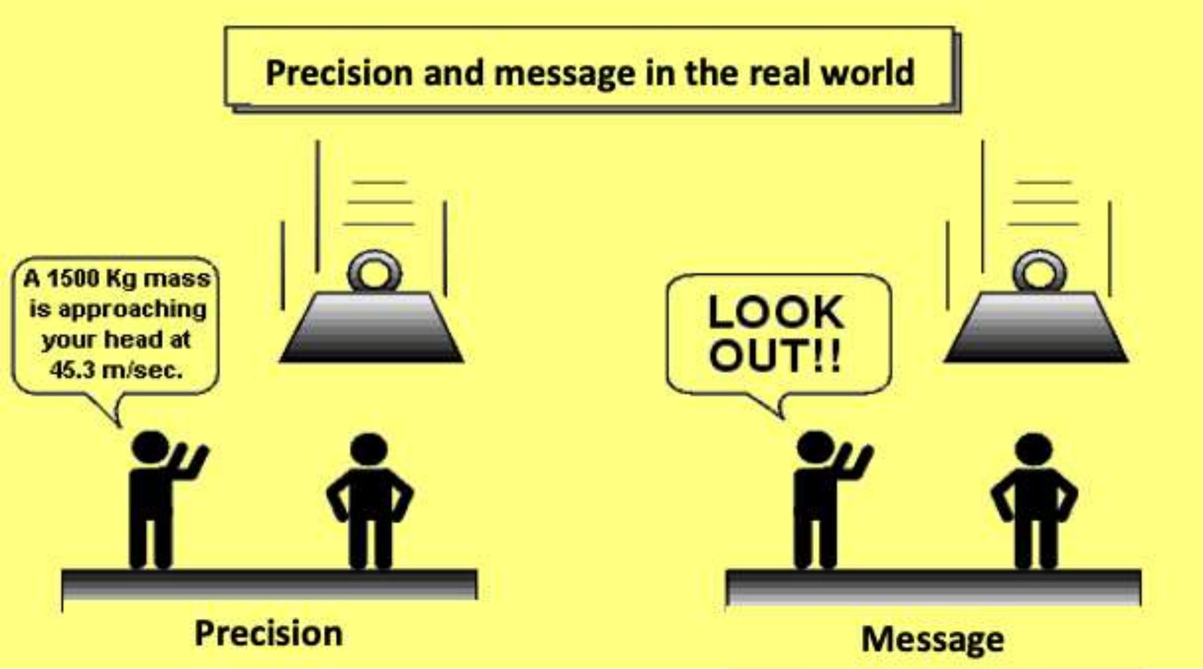
Amenable Deaths per 100,000 of population, 2015.
Downloaded from https://ec.europa.eu/eurostat/statistics-explained/pdfscache/41683.pdf,
1 July 2021, published by Eurostat

The chart shows that men (blue) suffer more unnecessary deaths than women (yellow), and that this happens in every country in Europe. If we look at the countries on the left of the chart, we see that the "bad" countries range from Poland through to Lithuania—and they are all ex-Eastern Europe. The message of the data, that men are the main problem, and the old Eastern Europe is the region where the problem is worst is clear—without numbers.

This chart is also a good example of the power of sorting the data by a relevant factor, in this case by total amendable deaths per 100,000 of population. The countries on the right are not part of the EU, they provide additional information, but they are not distracting the reader. If I was being super picky, I would have started the Y-axis on 0, but it does not materially change the message.

8. Convey the Message Not Precision

The previous point and the amenable deaths chart are also examples of the next tip, convey the message not the precision. The image below is a variation on a popular meme, making the point that we need to communicate the message (look out), as opposed to precision (e.g., the exact weight and speed of the falling object).


Precision and message in the real world
A 1500 Kg mass is approaching your head at 45.3 m/sec.
LOOK OUT!!
Precision
Message

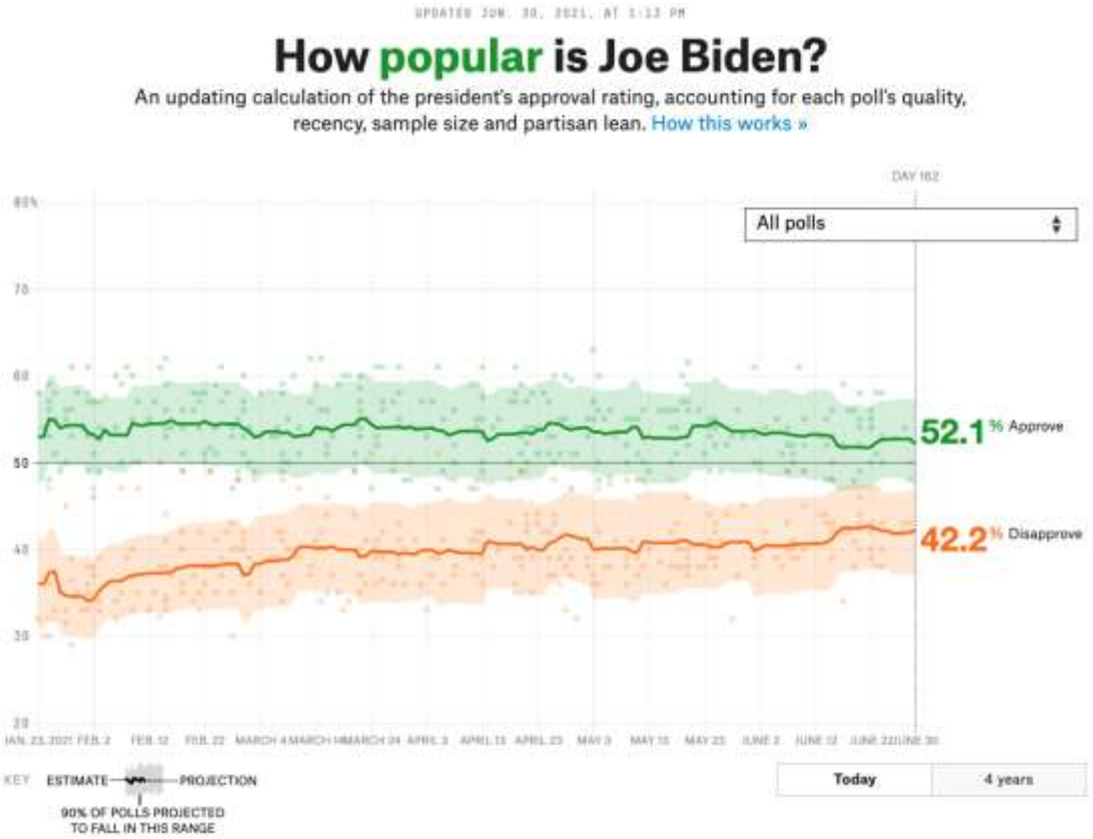
In terms of complex information, the message is often contained with just a few factors, clusters, or variables. In these cases, focus on the key elements, relegating the supporting details to a follow-up or appendix. In a paper you might want to publish all the relevant statistics and residuals, but in your presentation focus on the message.

9. Don't Bury the Story

Too many presentations can look like a page from a Where's Waldo book (or Where's Wally in other countries). If you show too much at once, the message is harder for the audience to detect. If you have a line chart with, say, seven lines on it then consider one of the following:

- Use builds to let the story unfold. Perhaps show one line, then sequentially add the other lines, commenting each time why this line is there and what the message is.
- Show several charts, each with, say, just two lines. Look back to the FiveThirtyEight example of President Biden's approval rating further up this paper. Many market researchers would have shown all seven lines (President Biden and the six other Presidents) on a single chart. But, by showing six charts, the cognitive load on the reader is massively reduced.
- Reduce the number of lines. Do you really need all seven lines for the message you are communicating? In some cases, the seven lines will be six named brands and the seventh will be "Other"—where other is different things for different people and the charting of it adds very little.

FiveThirtyEight provides another good illustration of how to avoid burying the story in their President Biden approval tracker, shown below, as of 1 July 2021.



UPDATED JUN. 30, 2021, AT 1:12 PM

How popular is Joe Biden?

An updating calculation of the president's approval rating, accounting for each poll's quality, recency, sample size and partisan lean. How this works »

The dark green and dark orange lines show the key averages, the lighter colored regions show 90% of the variation in individual polls (and help illustrate if there is a regular overlap of Approve and Disapprove). The faint green and orange dots are there for anybody who wants a more detailed picture, but they are not part of the message and they do not get between the reader and the message.

10. Dale Carnegie in Modern Form

Dale Carnegie's famous advice on presenting was "Tell the audience what you're going to say, say it; then tell them what you've said." This advice remains relevant today, but it is important to put it in the context of a modern presentation. The aim of repetition is not to simply say the same thing three times, but to convey the same message with three legs to the stool of understanding.

The first time you tell them, you should convey the core point of the message, for example, "We have identified that by doing X and Y you can create a better estimate of Z." Then you are going to provide evidence for your message—this is usually the bulk of the presentation. Then you are going to explain how to use the message. This is the same PEE that is taught to high school students—Point, Evidence, Explain.

One Tip to Rule Them All

At all times remember you are not creating your presentation for you to read, so don't design the presentation to be exactly the sort of presentation you like to receive. When you are creating a presentation, you already know the story, you already know why you are including various elements, but this is unlikely to be true for your audience. Your preferences in terms of tables versus charts, detail versus big picture are not relevant to the presentation you create. What is relevant is what works for your audience, i.e., what is effective.

Summary

If you don't effectively communicate your message, all your work, no matter how brilliant, is likely to be wasted. You need to focus on what works for your audience and create something that is going to be understood and hopefully acted upon.

These ten tips are going to help you be more effective, but only if you really focus on what works for your audience, in terms of your message, this time.

Ray Poynter

IMPROVING ACCESSIBILITY FOR ONLINE SURVEYS: LOOKING INTO INCREASING THE ONLINE SURVEY EXPERIENCE FOR INDIVIDUALS WITH VISION IMPAIRMENTS

NATHAN WIGGIN
COMENGAGE

INTRODUCTION

People with disabilities are both visible and invisible. According to a 2014 study conducted by the US Census Bureau, approximately 27% (or 85.3 million) of Americans have some type of disability (18% or 55.2 million have a severe disability). This includes:

- 12.3 million people who have severe difficulty seeing, including 1.6 million adults with full blindness,
- 17.1 million people with a serious hearing impairment, including 3.4 million who are deaf, and
- 48.2 million people with a functional (physical) limitation.

Worldwide, it is estimated that upwards of 190 million people experience significant disabilities.

Considering accessibility research organizations can unintentionally exclude ability to participate in online research. Some examples:

- Low vision, color blindness, or full blindness could mean relying on a screen magnifier or screen reader,
- Limited dexterity can make it difficult to use a keyboard or a mouse, and
- People with hearing impairments may need to rely on transcripts or captions for media content.

This is a large group of people who often get overlooked, especially in survey design. Unless a researcher is specifically targeting research toward one of these groups, we often put in minimal, if any, thought on the usability of the online survey experience for people with disabilities.

Not only does this invoke frustration, but it also reduces representation of an important part of the population, as, like all of us, most respondents would rather quit than deal with a difficult, or impossible, survey.

Moreover, not taking accessibility into account could lead to legal challenges. In 2018, the case United States vs. Astria Health determined that websites must be just as accessible to people with disabilities as they are to able-bodied individuals.

Additionally, including people with disabilities in research provides greater opportunities to learn from consumers who spend the same money as those without a disability. Do researchers and platforms want to ignore this large segment of the population and miss out on this hidden market?

This paper provides a case study for a survey conducted for a regional transit agency in Washington State. It outlines the journey from realizing that accessibility needed to be addressed to several survey iterations to address the problem. It concludes with some insight into the future direction of accessibility design as well as key takeaways for the reader to consider when designing a survey.

The primary objective of this paper is to raise awareness among researchers so that you have a basic understanding of accessible-friendly study design and take these issues into consideration when designing your survey.

This paper closely follows the presentation given at the Sawtooth Software Conference in April 2021 at San Antonio. There are links to three sample survey iterations that take the reader along. It is highly recommended that, to get the full experience, the reader either watch the presentation or follow through the examples provided.

PACKING FOR THE JOURNEY

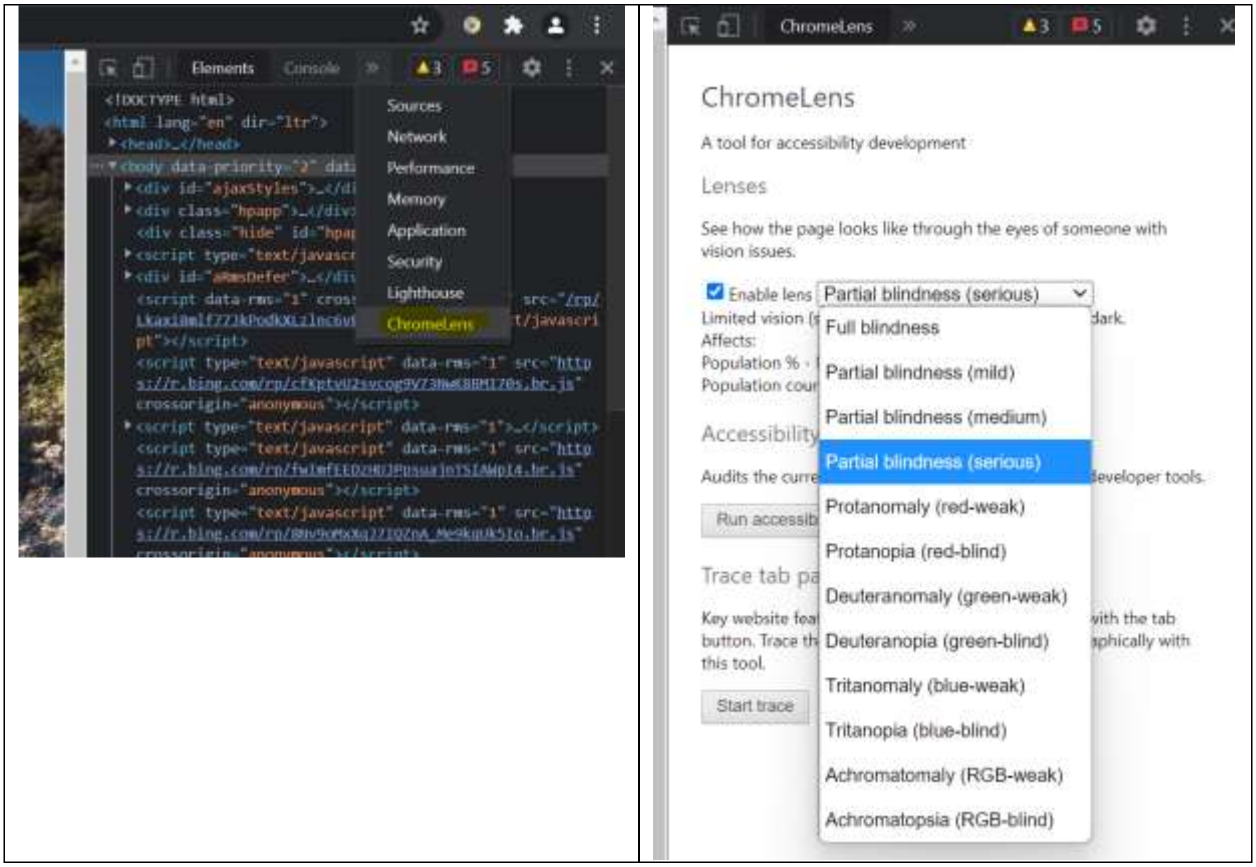
Some tools are needed to fully understand the difficulties of taking a survey that is not designed with accessibility. We recommend acquiring the following tools:

- JAWS Reader by Freedom Scientific (free version is available)
- ChromeLens, an extension for Google Chrome (free)
- Headphones

JAWS is a simple download and install. The free version works in 30-minute intervals before the computer needs to be reset. For the purposes of this paper that should be plenty of time.

ChromeLens is an extension for Google Chrome. Once installed, press F12 which will bring up the developer tools. Click the expander button and select ChromeLens to bring up the extension.

ChromeLens provides a variety of options for simulating visual impairment as well as tools to run accessibility testing and see how webpages are understood by screen readers. We recommend enabling the lens at either the medium or serious partial blindness setting.



```
<!DOCTYPE html>
<html lang="en" dir="ltr">
 ▸<head>…</head>
…▾<body data-priority="2" data
   ▸<div id="ajaxStyles">…</di
   ▸<div class="hpapp">…</div
    <div class="hide" id="hpa
   ▸<script type="text/javascr
   ▸<div id="aRmsDefer">…</di
    <script data-rms="1" cros                         src="/rp/
    LkaxI0mlf77JkPodkXLzlnc6v                    t/javascri
    pt"></script>
    <script type="text/javascript" data-rms="1" src="http
    s://r.bing.com/rp/cfKptvU2svcog9Y73Hw8BM170s.br.js"
    crossorigin="anonymous"></script>
   ▸<script type="text/javascript" data-rms="1">…</script>
    <script type="text/javascript" data-rms="1" src="http
    s://r.bing.com/rp/fwlmfEED2HDJPpsuajnTSIAWp14.br.js"
    crossorigin="anonymous"></script>
    <script type="text/javascript" data-rms="1" src="http
    s://r.bing.com/rp/8Nv9oMxXq77IQZnA_Me9kqUk5Io.br.js"
    crossorigin="anonymous"></script>
```

ChromeLens

A tool for accessibility development

Lenses

See how the page looks like through the eyes of someone with vision issues.

☑ Enable lens Partial blindness (serious)

Limited vision (s   dark.
Affects:
Population % ·
Population cour

Accessibility

Audits the curre   developer tools.

Run accessib

Trace tab pa

Key website fea   vith the tab
button. Trace th   aphically with
this tool.

Start trace

Of note, individuals who need to use screen readers do NOT navigate using keyboard and mouse. All computer interaction is done via a series of hot keys. There are schools for the blind where this information is taught. It is not intuitive for most people so readers will most likely experience frustration while learning to navigate using screen readers. This is true even on well-designed websites.

Take 1: No Accessibility Considerations

The first iteration of the survey was designed using Sawtooth Software's Lighthouse Studio 9.9. Additionally, no considerations were taken toward accessibility. This was the true "out of the box" experience. All survey iterations in this paper should take less than 3 minutes to complete. They consist of very few questions and are simply demonstrations of accessibility access and issues.

We recommend that readers attempt to complete each iteration two times. The first run-through should be a quick run-through in a typical manner using mouse and keyboard. The second run-through should be done using the screen reading software and ChromeLens to mimic partial (or full) blindness.

To attempt iteration 1 click the following link to launch the survey in Chrome: First Attempt, No Considerations. Once the survey is up, launch the screen reader, navigate to ChromeLens, pick your visual impairment, and proceed.

With no considerations taken toward accessibility, the survey is impossible to complete. Respondents are unable to know what item(s) is selected and unable to navigate as there are no indicators for the screen reader to use.

Immediately upon client testing, this survey was called a failed effort and we either had to address/fix the issue or we would lose the contract as an important part of our sample was individuals with a visual impairment.

Take 2: Separate but Equal

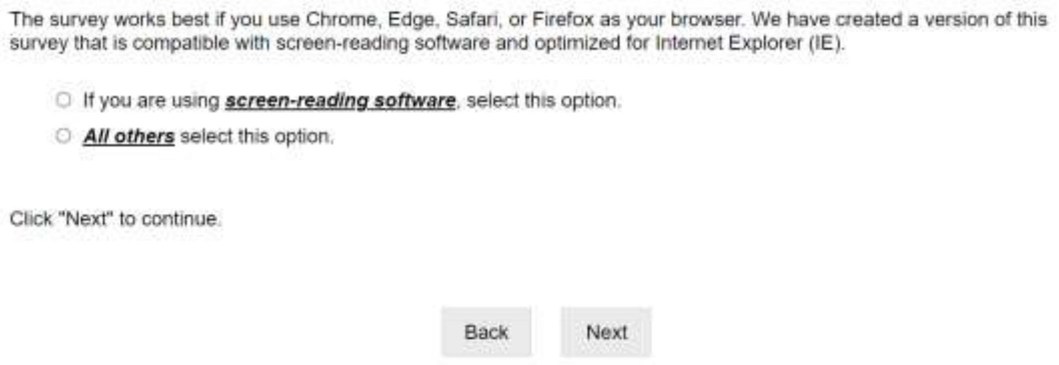
Realizing that the initial study design would not work, we began working on how to resolve the issues. The second design used a bifurcated approach. Two surveys were programmed. The first was designed to be visually appealing using expanding grids and standard question types. The second was specifically designed to work with screen readers. This is what is known as "Separate but Equal."

There are two definitions for Separate but Equal that apply to our situation. The practical definition is creating separate items (or experiences) for separate individuals. The legal definition is "The doctrine set forth by the U.S. Supreme Court that sanctioned the segregation of individuals by race in separate but equal facilities but that was invalidated as unconstitutional" (Merriam-Webster dictionary).

A good example of separate but equal is from the civil rights era where facilities were segregated such as drinking fountains or restrooms in the 1950s. While our goal was to provide the best quality experience for all users (unlike in the 1950s), the way in which this was accomplished is generally frowned upon and there are numerous issues resulting from its use.

Open the following link in chrome to attempt the second iteration: Second Attempt: Separate but Equal.

Launch the screen reader and ChromeLens, however run through the survey prior to enabling a vision impairment. The first thing you will notice is a question asking if you are using a screen reader. This is an obvious indicator that there are separate surveys.



The survey works best if you use Chrome, Edge, Safari, or Firefox as your browser. We have created a version of this survey that is compatible with screen-reading software and optimized for Internet Explorer (IE).

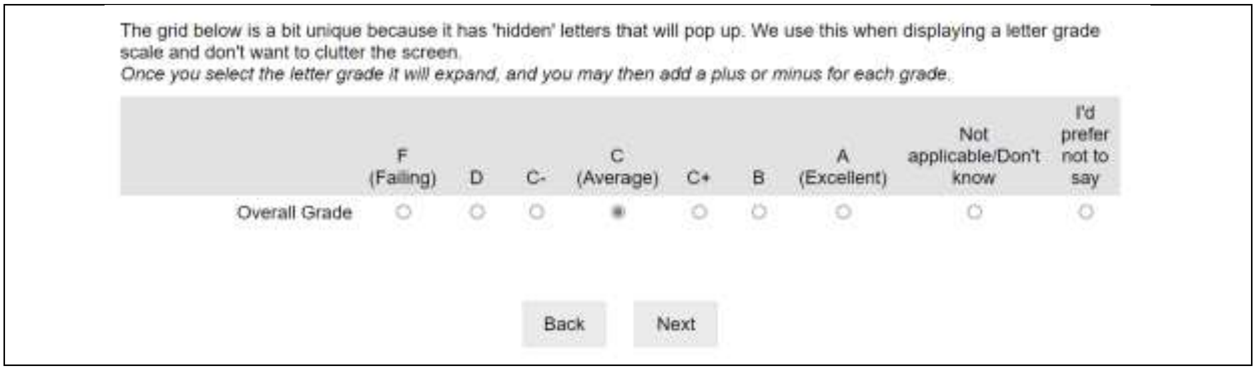
○ If you are using ***screen-reading software***, select this option.

○ ***All others*** select this option.

Click "Next" to continue.

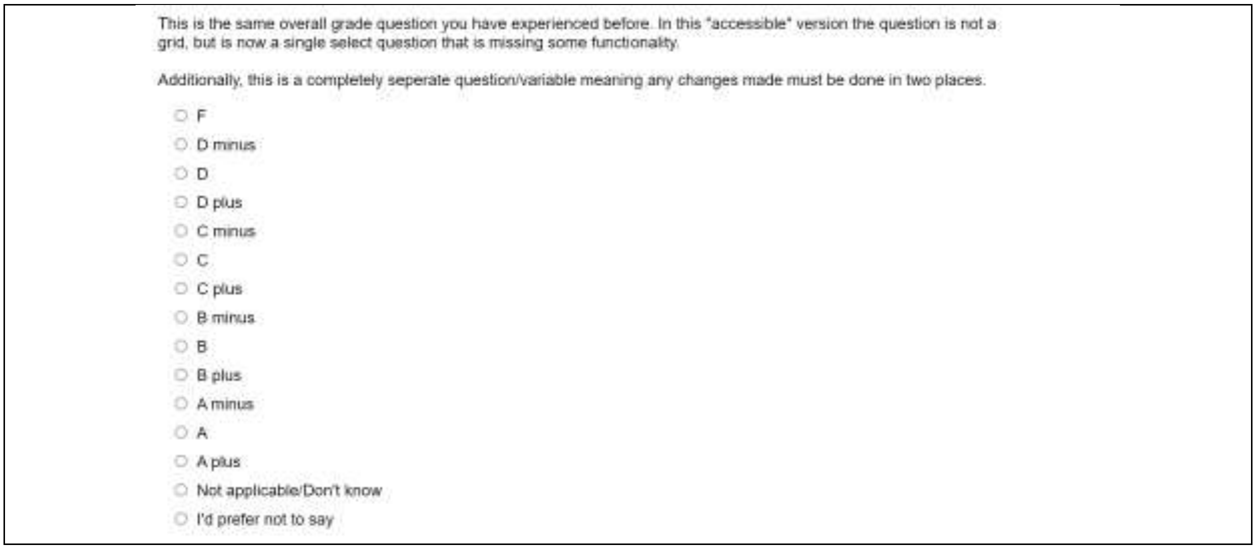
The differences become more apparent as the survey progresses. The visual version has the expanding grid question, and the low vision version just lists the question as a single select. These are, in fact, separate questions. They are programmed separately and write to different variables in the data file.

Attempting the visual version using a screen reader results in failure, again as a result of not being able to navigate. Issues persist in the accessible version as well. For example, on the screen pictured below the screen reader indicates there are 51 possible selection fields. There is no tag that indicates one group of response options as belonging to one question and the next group to another question. Therefore the software just totals all selectable items. While this did satisfy the basic demands of our client and we did go to field, they were not happy with the overall quality as it was easy to tell there were two surveys, and there were still several accessibility issues to address.



The grid below is a bit unique because it has 'hidden' letters that will pop up. We use this when displaying a letter grade scale and don't want to clutter the screen.
Once you select the letter grade it will expand, and you may then add a plus or minus for each grade.

	F (Failing)	D	C-	C (Average)	C+	B	A (Excellent)	Not applicable/Don't know	I'd prefer not to say
Overall Grade	○	○	○	●	○	○	○	○	○

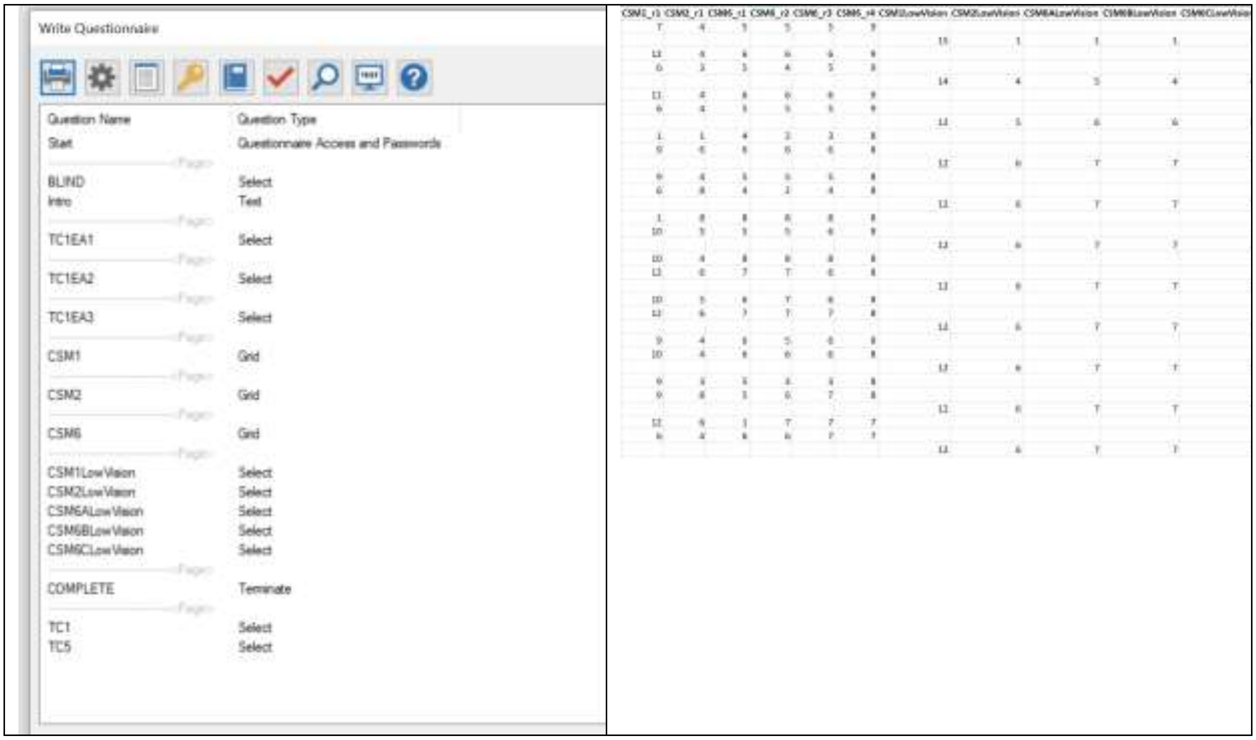


This is the same overall grade question you have experienced before. In this "accessible" version the question is not a grid, but is now a single select question that is missing some functionality.

Additionally, this is a completely seperate question/variable meaning any changes made must be done in two places.

○ F
○ D minus
○ D
○ D plus
○ C minus
○ C
○ C plus
○ B minus
○ B
○ B plus
○ A minus
○ A
○ A plus
○ Not applicable/Don't know
○ I'd prefer not to say

In addition to usability concerns, the programming created large quality control and data analysis difficulties. Each question needs to be programmed twice, once for the normal survey and a second time for the accessible survey. Therefore, all changes needed to be made in multiple places resulting in errors of omission for one version or the other. Additionally, the split approach created a complex datafile. Variables had to be recoded and merged, again resulting in an increased workload and possibility of errors.



Question Name	Question Type
Start	Questionnaire Access and Passwords
Page	
BLIND	Select
Intro	Text
Page	
TC1EA1	Select
Page	
TC1EA2	Select
Page	
TC1EA3	Select
Page	
CSM1	Grid
Page	
CSM2	Grid
Page	
CSM6	Grid
Page	
CSM1LowVision	Select
CSM2LowVision	Select
CSM6ALowVision	Select
CSM6BLowVision	Select
CSM6CLowVision	Select
Page	
COMPLETE	Terminate
Page	
TC1	Select
TC5	Select

CSM1_r1	CSM2_r1	CSM6_r1	CSM6_r2	CSM6_r3	CSM6_r4	CSM1LowVision	CSM2LowVision	CSM6ALowVision	CSM6BLowVision	CSM6CLowVisio
7	4	5	5	5	9					
						15	1	1	1	
12	4	6	6	6	9					
6	3	5	4	5	8					
						14	4	5	4	
13	4	8	6	8	9					
6	4	5	5	5	9					
						12	5	6	6	
1	1	4	2	2	8					
9	6	6	6	6	8					
						12	6	7	7	
6	4	5	5	5	8					
6	8	4	2	4	8					
						12	6	7	7	
1	8	8	8	8	8					
10	5	5	5	6	9					
						13	6	7	3	
10	4	8	8	8	8					
12	6	7	7	6	8					
						13	6	7	7	
10	5	6	7	6	8					
12	6	7	7	7	8					
						12	6	7	7	
9	4	8	5	6	8					
10	4	6	6	6	8					
						13	6	7	7	
6	5	5	4	5	8					
9	6	5	6	7	8					
						12	6	7	7	
12	6	1	7	7	7					
6	4	6	6	7	7					
						12	6	7	7	

TAKE 3: THIRD TIME IS THE CHARM (SORT OF)

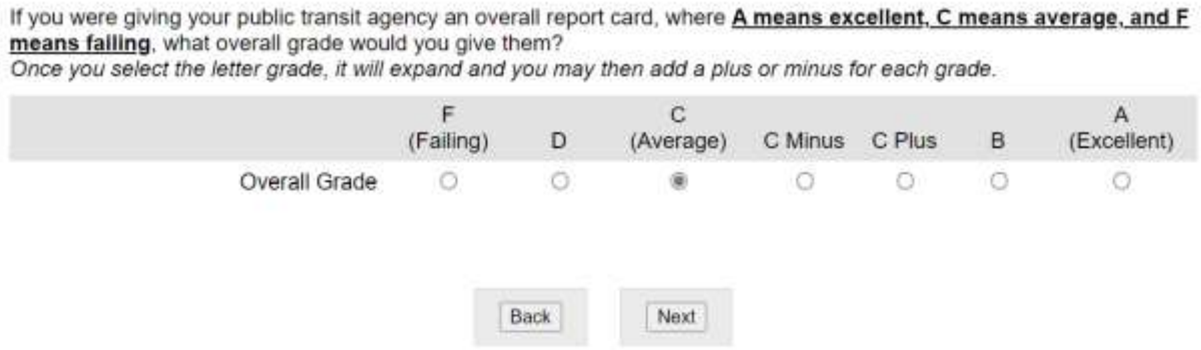
The study for which this survey was designed was an ongoing study consisting of multiple waves of data collection. Each wave was about 6 to 9 months apart. We spent the time after Take 2 working on improving the survey experience further. Our goal was straightforward, create a single, usable experience for all respondents.

Thanks to the diligent efforts of Sawtooth Software Support we managed to piece together a fairly solid survey.

Open the following link in Chrome to attempt the third, and most recent iteration: Take 3.

For this iteration we recommend setting ChromeLens either on severe or full blindness and relying on the screen reader for the first attempt. This iteration is much cleaner. It is a single, fluid survey that incorporates all of the visual elements, such as the expanding grid, while still being accessible.

While testing, we made an important discovery. Screen readers read left to right, top to bottom. They do not go backward to see if something above (or to the left) it has changed. Therefore, we needed to modify the way the expanding grids worked. In the first two iterations, when a letter grade was clicked, the grid would expand to the left (indicating a "minus" grade) and the right (indicating the "plus" grade). This was redesigned so all expansion happened on the right side of the clicked button. Thus, the screen reader was able to pick up the expanding grid. Additionally, we originally used the mathematical symbols for plus (+) and minus (-). However, when testing we found that the screen reader doesn't have context and would simply read "c dash" which could be confusing to respondents. Therefore, we wrote out the words for clarity.

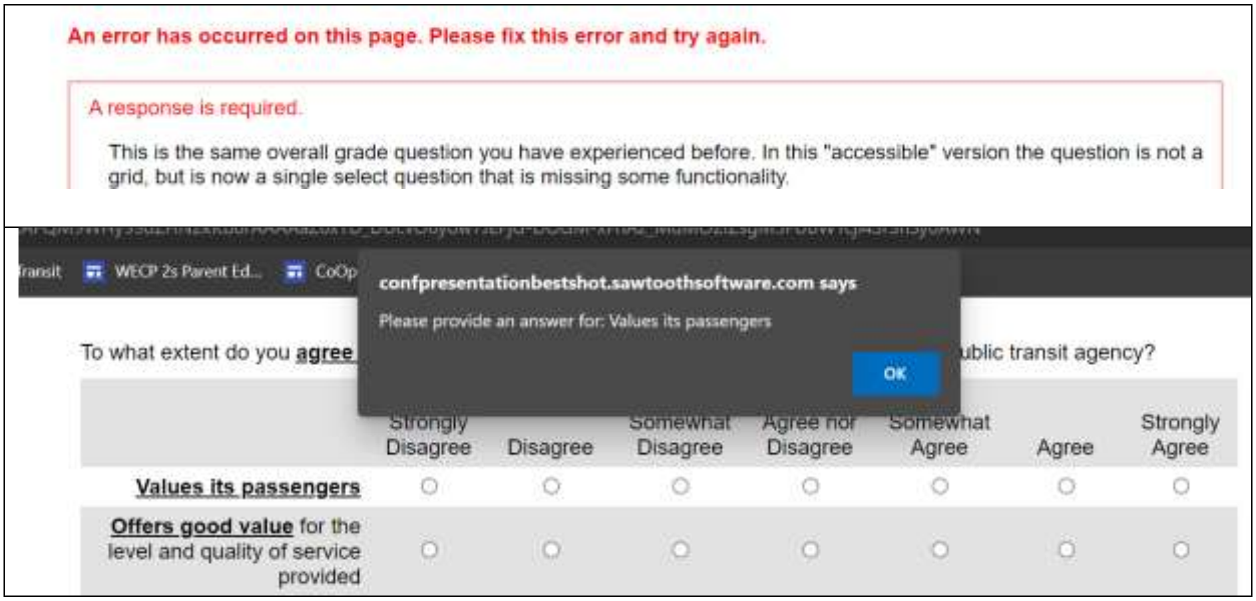


If you were giving your public transit agency an overall report card, where **A means excellent, C means average, and F means failing**, what overall grade would you give them?
Once you select the letter grade, it will expand and you may then add a plus or minus for each grade.

	F (Failing)	D	C (Average)	C Minus	C Plus	B	A (Excellent)
Overall Grade	○	○	●	○	○	○	○

We also realized that the normal behavior for error messages is to add red text to the top of the screen. Something needed to be done to address this as the screen readers did not pick up on the error messages unless the user reloaded the screen or started the screen reader over at the top. This resulted in several test participants being unable to advance, not knowing why, and closing the survey.

Our solution was a simple change to the JavaScript code used for error messages; we added a command that changed the error message to an Alert which caused a pop up stating the error. The screen readers picked up and read the error message and respondents were able to find the missed question and continue the survey.



An error has occurred on this page. Please fix this error and try again.

A response is required.

This is the same overall grade question you have experienced before. In this "accessible" version the question is not a grid, but is now a single select question that is missing some functionality.

confpresentationbestshot.sawtoothsoftware.com says

Please provide an answer for: Values its passengers

To what extent do you **agree** … ublic transit agency?

	Strongly Disagree	Disagree	Somewhat Disagree	Agree nor Disagree	Somewhat Agree	Agree	Strongly Agree
Values its passengers	○	○	○	○	○	○	○
Offers good value for the level and quality of service provided	○	○	○	○	○	○	○

Continuing Refinements

The third iteration was used in October 2020 and since that time Sawtooth Software has worked diligently to improve the ground up experience and with the release of Lighthouse 9.11.0 in April 2021 native accessibility has improved dramatically.

COVID-19 has continued to have an impact on this study so we have not incorporated all of the improvements from Lighthouse 9.11.0 and beyond. However, the next wave is currently scheduled for October 2021 and will be redesigned from the ground up using the latest version. While we are excited to incorporate the improvements, there are additional client asks. The fall 2021 iteration will be conducted in 12 languages and we have been working with Sawtooth Software support to create a single survey with one set of variables that is not only accessible, but also conducted in multiple languages. We were able to do a test run of the language portion in July 2021 and we look forward to the challenges of combining all elements together.

Conclusions and Takeaways

Our team has learned a lot while working on this project and we would like to conclude by reiterating the most important takeaways for the reader:

1. *Consider the abilities and needs of others*: As a species, we take vision for granted. The vast majority of us can see well enough to use a computer. We do not think about how an impairment, be it vision or dexterity, impacts what we do day-to-day. As researchers, this is something that we need to get into the habit of doing. At a minimum, we need to think and consider how experiences may differ for individuals with differing abilities.
2. *Try this yourself*: In the presentation, I walked through the three examples rather quickly, and while I provided the links, a conference setting is not the best place to blindfold oneself and attempt a survey. But now *is* the time. The tools are easily accessible. Please

go through the links provided and try this yourself, and take a few minutes next time you program a survey to test it for accessibility. It is a good habit to get into.

3. *Get an expert*: This is probably not needed for every study, but if you find yourself in a situation where accessibility is of high importance, we recommend finding an accessibility expert. They are invaluable.

Nathan Wiggin

CRITICAL USER JOURNEY MAD LIBS— A NEW TECHNIQUE FOR CONCEPT DEVELOPMENT

JON GASS
PINTEREST
AIDEEN STRONGE
GOOGLE

ABSTRACT

CUJ Mad Libs is a technique used to better define User needs. A common step in Product Development is to gather feedback on wireframes, but we may unknowingly have insufficient information to fairly represent and compare scenarios, leading to innovative ideas being cut—this paper proposes an alternative approach.

INTRODUCTION

CUJ (Critical User Journey) Mad Libs is a lightweight technique that is used to better understand nuances around user requirements. The technique was first developed at Google in 2020 by the first author, and is particularly useful in complex spaces in which features are more likely to transition from product specifications to interaction designs before there is opportunity to gather customer insights. While the technique can be used as a standalone process, the technique has two additional benefits when combined with other user experience (UX) processes.

1. It helps researchers represent user scenarios more accurately in MaxDiff surveys or concept studies, facilitating the development of innovative solutions to more precisely defined problems.
2. It draws cross-functional partners into the process of creating Critical User Journeys (CUJs) which can sometimes feel like a solitary product excellence effort undertaken only by user experience professionals.

HOW MAD LIBS WORKS

Individuals complete text-based templates for problems associated with their most crucial CUJs. By iteratively refining these problem statements, user researchers can get faster alignment with stakeholders on which problems to focus on. The method is best used during planning phases but can also be used as a precursor to benchmarking or wireframe evaluations.

CUJ Mad Libs can build out CUJs with the input of customers before design assets are created. While concept studies using wireframes are often used for evaluating concepts, new opportunities may be missed or even misrepresented because insufficient time has been spent refining the problem statements before committing to design.

Types of questions you can answer with this method

- Why are certain problems considered to be more important to solve than others?
- Which feature team ideas are worthy of exploring further?
- What are the CUJs associated with a customer complaint?
- What expectations do users have about how these problems are solved?
- What are anticipated barriers to solutions being adopted?

BACKGROUND

To illustrate the gap that CUJ Mad Libs fill, let us first consider a common sequence of research techniques that might be leveraged during the product development cycle. In learning about how customers are interacting with or feeling about our products we might leverage data science or customer satisfaction surveys (CSAT); and to better understand the reasoning behind these behaviors and satisfaction ratings we might use contextual surveys, customer interviews or diary studies. After mining raw text responses from these methods, solutions to pain points might be prioritized with a MaxDiff survey, but what if these solutions did not resonate with customers because the user scenarios associated with the problems are only partially understood (see Figure 1)? A similar misstep could also occur if a product team rushed into wireframing solutions for a concept study too early. CUJ Mad Libs addresses this gap (see Figure 2). By acknowledging that customers have different vantage points of a problem, a problem statement can be iterated on so that it can be more accurately prioritized in subsequent rounds of research. Further, one can begin a MaxDiff study with greater confidence that the most important ideas will shine.

Figure 1: Illustrating a Potential Pitfall in a Common Sequence of User Research Techniques

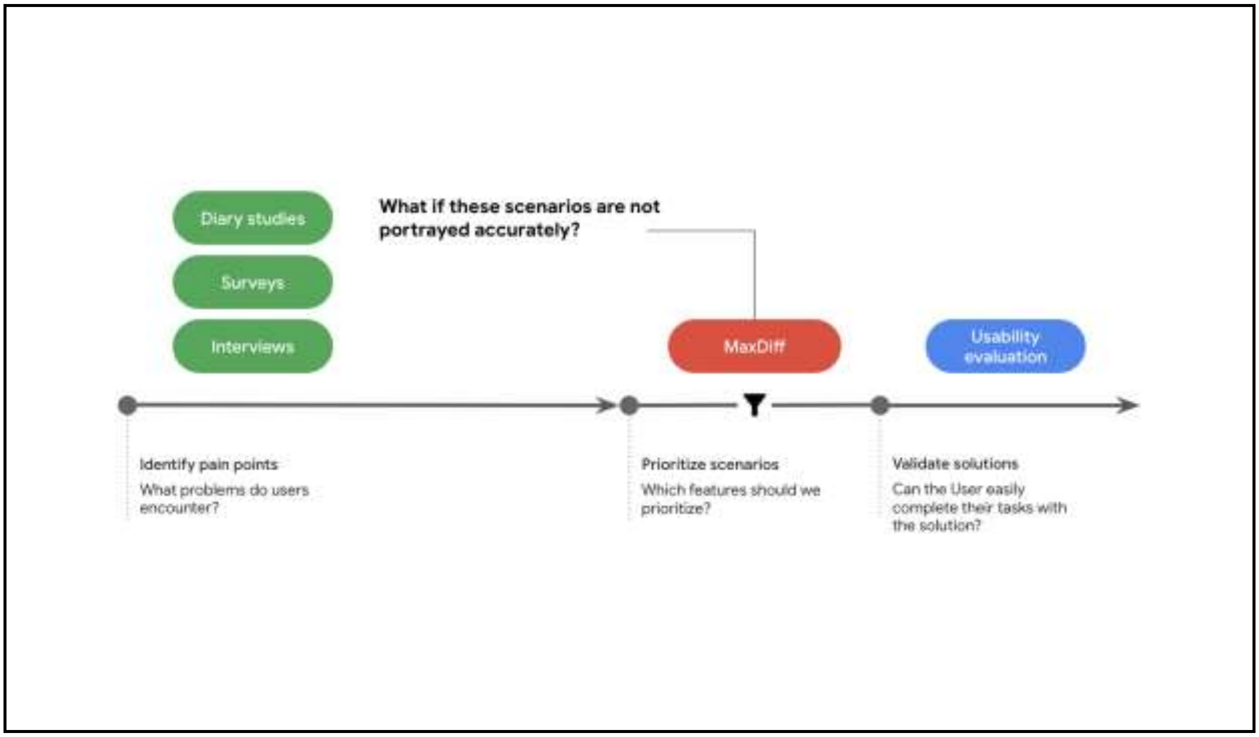
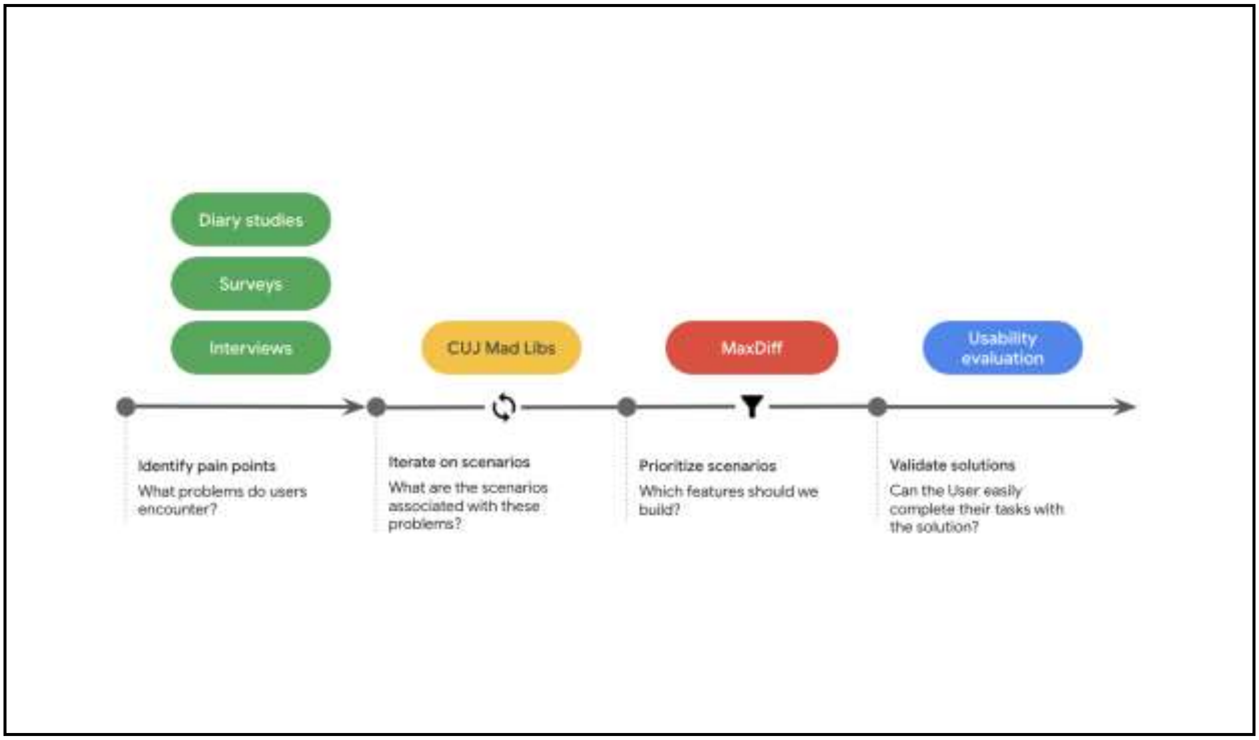

Diary studies
Surveys
Interviews
What if these scenarios are not portrayed accurately?
MaxDiff
Usability evaluation
Identify pain points
What problems do users encounter?
Prioritize scenarios
Which features should we prioritize?
Validate solutions
Can the User easily complete their tasks with the solution?

Figure 2: Illustrating Where the CUJ Mad Libs Process Occurs


Diary studies
Surveys
Interviews
CUJ Mad Libs
MaxDiff
Usability evaluation
Identify pain points
What problems do users encounter?
Iterate on scenarios
What are the scenarios associated with these problems?
Prioritize scenarios
Which features should we build?
Validate solutions
Can the User easily complete their tasks with the solution?

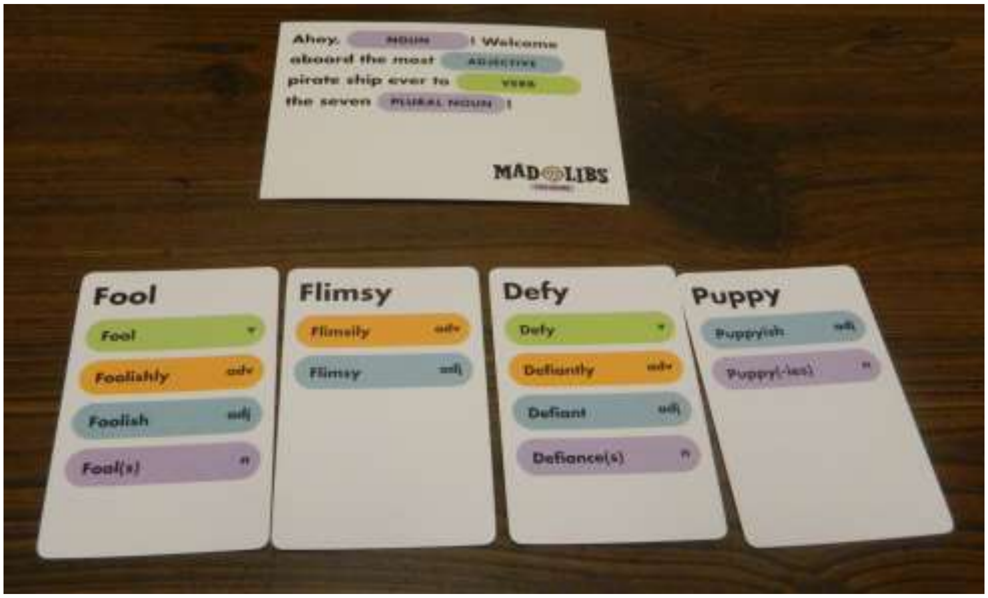
INSPIRATION FOR THE METHOD

The development of the method came about from being tasked with helping my team develop a roadmap in a complex space. In particular the team needed to get a deeper understanding of the problems and to prioritize them. We had gathered some pain points but concept testing was not feasible because we were not sure how these pain points occurred in the context of their tasks, and even if we could there were too many for this approach. Net, the team needed another way to arrive at a subset of problems to explore without designs.

Of all the unexpected places, inspiration came from a game called Mad Lib Cards. Mad Libs Cards is a template-based word game where players complete a sentence using words from a set of playing cards.

In the example below (Figure 3), the pre-filled sentence would read: "Ahoy (fool)! Welcome aboard the most (flimsy) pirate ship ever to (defy) the seven (puppies)!" Players read their sentences aloud and the player whose story receives the most votes from the other players wins the round.

Figure 3: Mad Libs Card Game

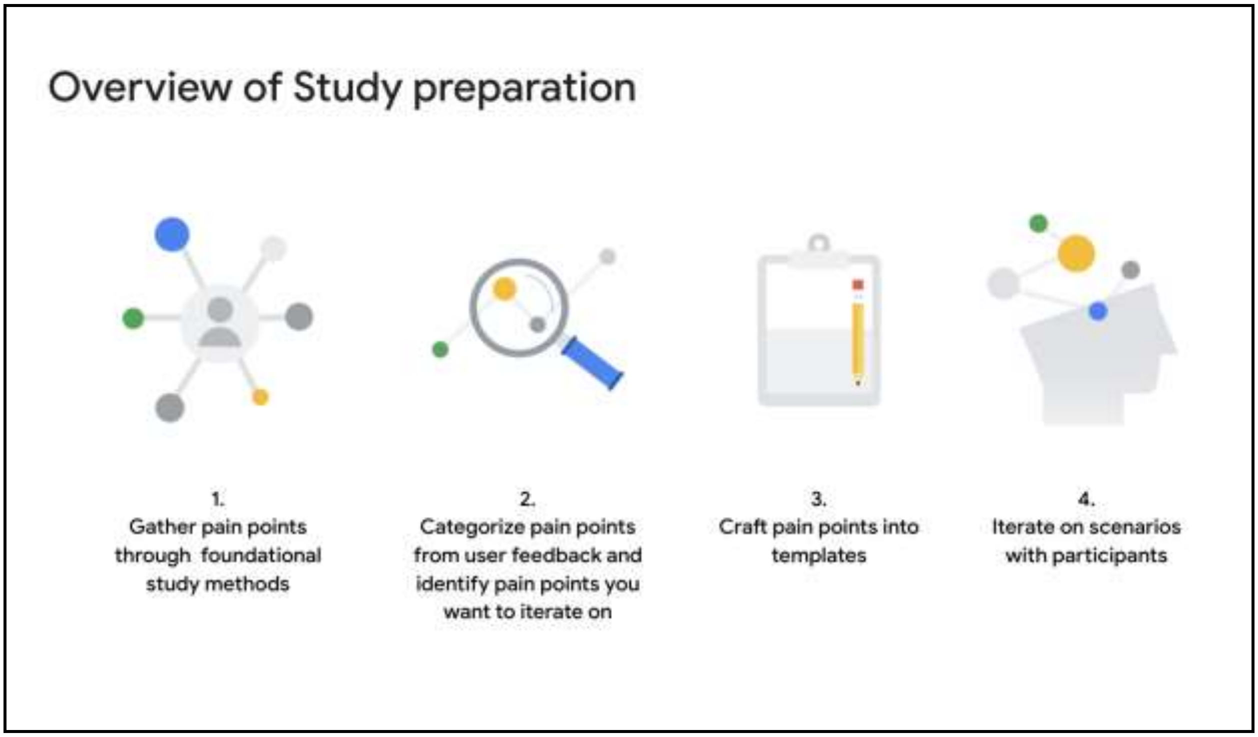


METHOD

Overview

At a high-level, there are four phases involved in the CUJ Mad Libs technique, which can optionally be followed by a MaxDiff survey. The first phase is to gather customer pain points from other sources such as surveys, customer support calls or even anecdotal insights from internal teams. Secondly, the pain points are categorized into themes of exploration, and an agreement is made among cross-functional stakeholders on the problem statements to focus on. Thirdly, these areas of focus are added to a template (shown in Figure 5b). And finally, the description of these scenarios are iterated on with the input of Users. In the sections below we go into further detail on phases 2, 3 and 4.

Figure 4: Illustrating Where the CUJ Mad Libs Process Occurs



Overview of Study preparation

1\.
Gather pain points through foundational study methods

2\.
Categorize pain points from user feedback and identify pain points you want to iterate on

3\.
Craft pain points into templates

4\.
Iterate on scenarios with participants

What to Expect

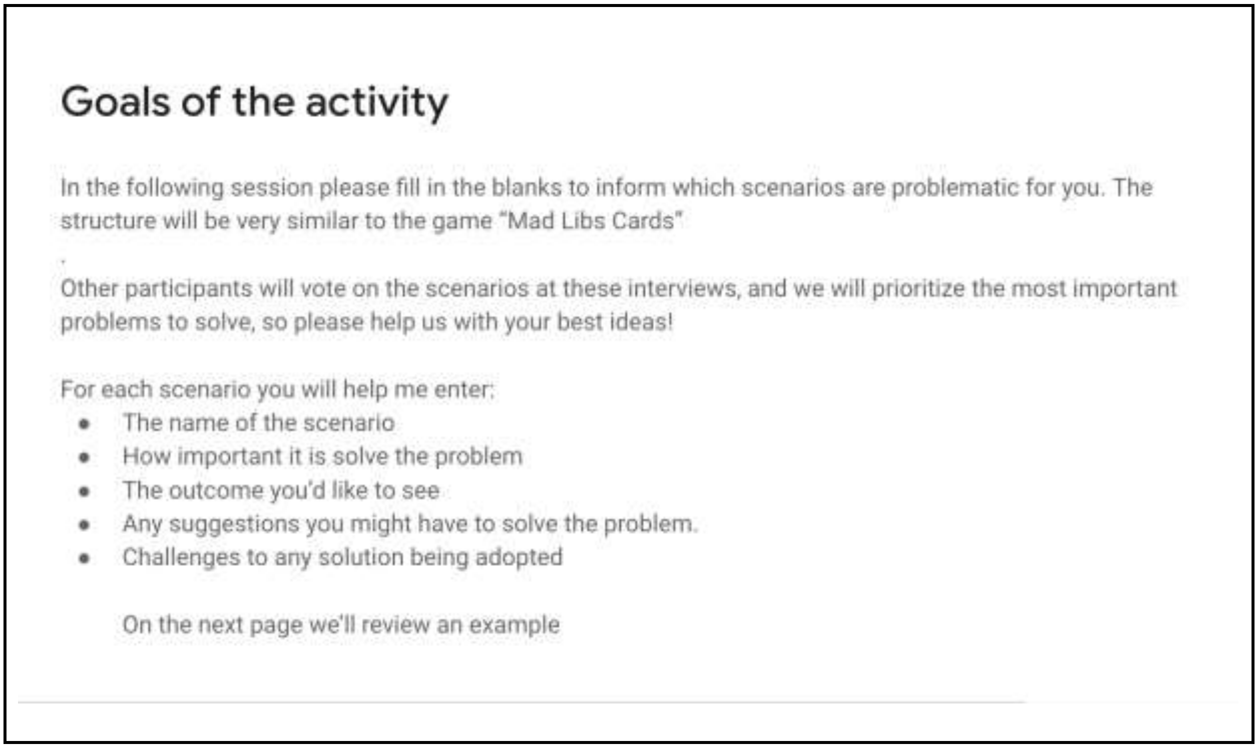
- Time Commitment: 30–60 mins per User
- Participants: 8–10 Users
- Planning Required: Researchers must have existing pain points to build scenarios around. This might be obtainable from surveys or previous field research. Assuming you have these, allow 1.5 weeks to build out the scenarios and get approval from stakeholders.
- Required materials:
  - Empty and pre-filled scenarios in a deck to present to the participant.
  - Additional document to take notes.
  - Survey responses from participants are helpful to jog their memories on pain points they want to construct CUJs for.
- Research Stage: Primary use: Foundational. Secondary use: Tactical Research
- Where does this fall in the product cycle?: MVP Product definition or Roadmap for v.2/v.x releases

How to Conduct a CUJ Mad Libs Study

1. Gather existing customer pain points from previous research.
2. Categorize pain points into themes you'd like to understand CUJs for.
3. Complete a template for each of these themes (see Figure 5b). Often you may only know the customer task name, or a pain point. Other times your team may already have a concrete idea of the outcome or solution ideas. Either situation is fine since you'll be inviting Users to fill in or reword the missing fields, and more importantly probing on the reasoning behind their answers (see Figure 5a for introductory text).
4. Interview at least 8 Users to help you complete your templates.
5. Iterate on problem statements between each session to make the problem statement precise and easily understandable. Occasionally customers might bring up a variant of the problem you are asking about. If this happens just clone the template, give it a different name and decide if it's worth asking others about.
6. Review the completed problem statements with your team, and decide on the subset of issues you'd like to prioritize using a technique like MaxDiff surveys.
7. Present the prioritized list of issues with the team. A typical example should include:
   - What resonated with Users,
   - Barriers to adoption,
   - Next steps or open questions. For example, pulling additional instrumentation data.
8. Once you have agreement on the subset of issues to tackle, wireframe out the CUJs and validate that the solutions meet User requirements.

Examples of Participant Materials

Figure 5a: Introduction to the Mad Libs Technique



Goals of the activity

In the following session please fill in the blanks to inform which scenarios are problematic for you. The structure will be very similar to the game "Mad Libs Cards"

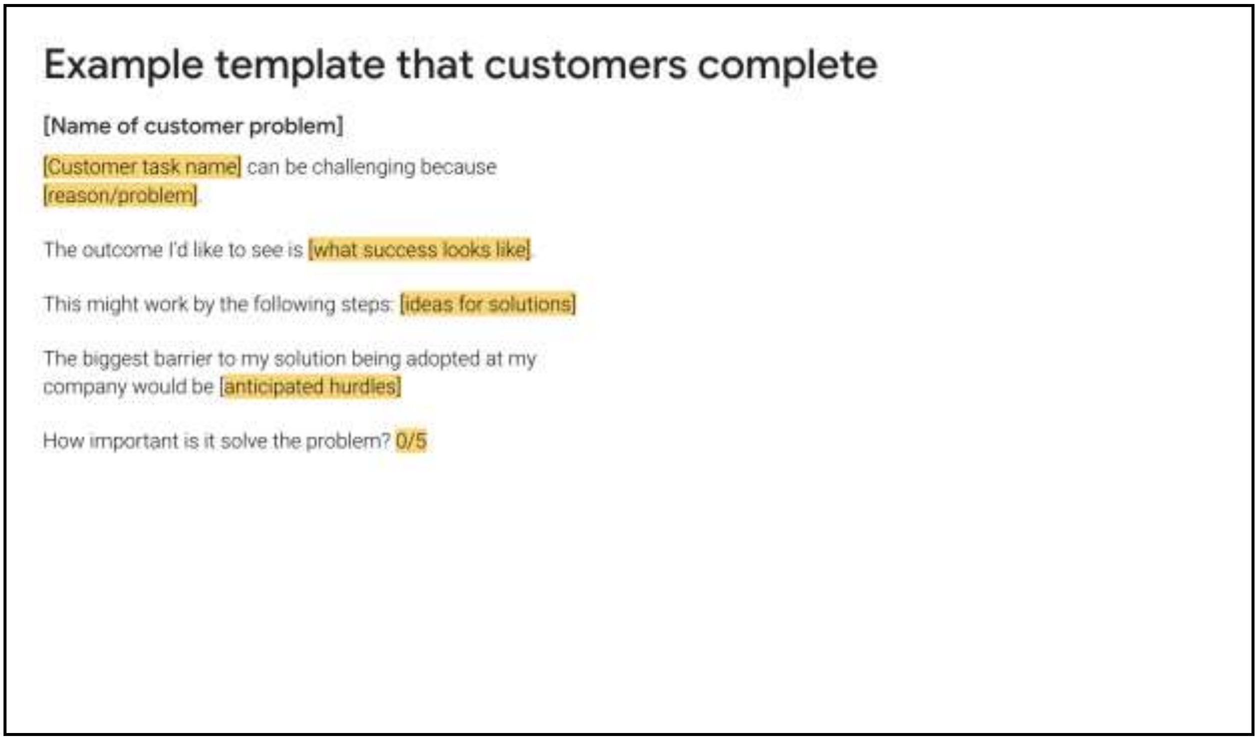
Other participants will vote on the scenarios at these interviews, and we will prioritize the most important problems to solve, so please help us with your best ideas!

For each scenario you will help me enter:

- The name of the scenario
- How important it is solve the problem
- The outcome you'd like to see
- Any suggestions you might have to solve the problem.
- Challenges to any solution being adopted

On the next page we'll review an example

Figure 5b: Example of Mad Libs Template



Example template that customers complete

[Name of customer problem]

[Customer task name] can be challenging because [reason/problem].

The outcome I'd like to see is [what success looks like].

This might work by the following steps: [ideas for solutions]

The biggest barrier to my solution being adopted at my company would be [anticipated hurdles]

How important is it solve the problem? 0/5

Figure 5b is an example of the template that Customers complete—there are six fields that a participant must complete or a Researcher might partially pre-fill. The key insights to gather include:

1. A name for the problem.
2. Why the task is tough.
3. The ideal outcome they are trying to achieve. This helps a product team identify opportunities that a participant might overlook or think is impossible.
4. Their ideas for a solution. Note this isn't to say that the team has to adopt it, but it opens up an opportunity for the Researcher to probe further on whether other experiences have led to their solution.
5. The barriers to the solution being adopted by others—this might draw out insights on workarounds to the problem from colleagues.
6. An importance rating for the problem—this provides a way to spot differences between segments of Users so that when you do send out your MaxDiff survey, you can be mindful of sending it out to a representative audience that you eventually segment.

Tips for conducting the interview with the customer:

1. Allow the customer to try some warm up CUJ Mad Libs.
2. Allow the customer to refer back to any previous pain points they may have submitted via screener surveys or diaries.
3. Allow each customer to refine the problem statement, so that you can update the templates with your team members between participants.
4. If a customer identifies a different pain point, create a new template for it.

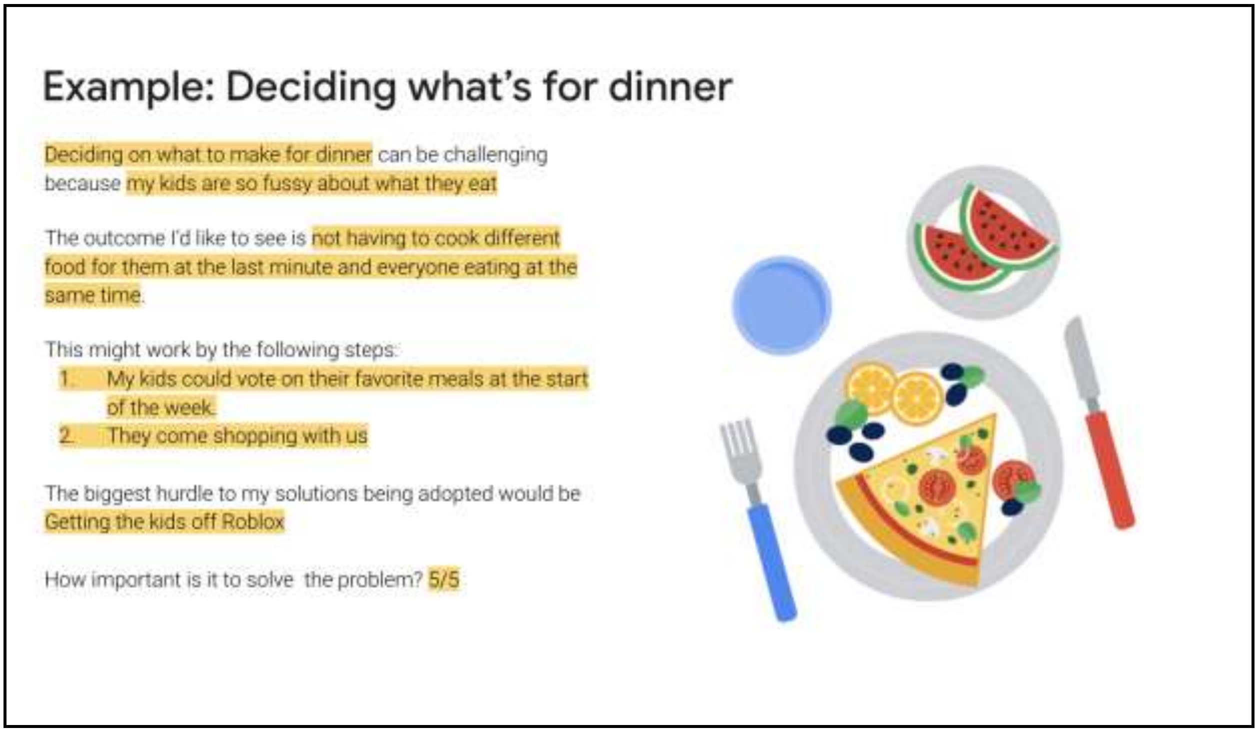
5. Ensure that participants understand the Scenario in the template: You will need for the problem statement to stand on its own in the MaxDiff survey, so make the necessary tweaks early.
6. Use the template to draw out more insights: If you have prefilled the template on a pain point that is new to the User, a great question to ask is “Can you think of a time that you encountered the problem? Tell me about it.” Recalling this will often help them recall more information.
7. Involve your XFN partners: identifying and pre-filling the template with partners not only aligns you, but it gives you a second set of eyes to identify problem statements that need to be consolidated with others.

Results: Examples

Due to reasons of confidentiality, results from previous studies cannot be shared, but Figures 6 and 7 illustrate two examples of outputs from different phases of the study.

Figure 6 illustrates how a completed template might look if a parent completed a CUJ Mad Lib for challenges with parenting. In this example, the participant task they have chosen is deciding on what to make for dinner.

Figure 6: Illustrating How a Completed CUJ Mad Libs Template Might Look



Example: Deciding what’s for dinner

Deciding on what to make for dinner can be challenging because my kids are so fussy about what they eat

The outcome I'd like to see is not having to cook different food for them at the last minute and everyone eating at the same time.

This might work by the following steps:

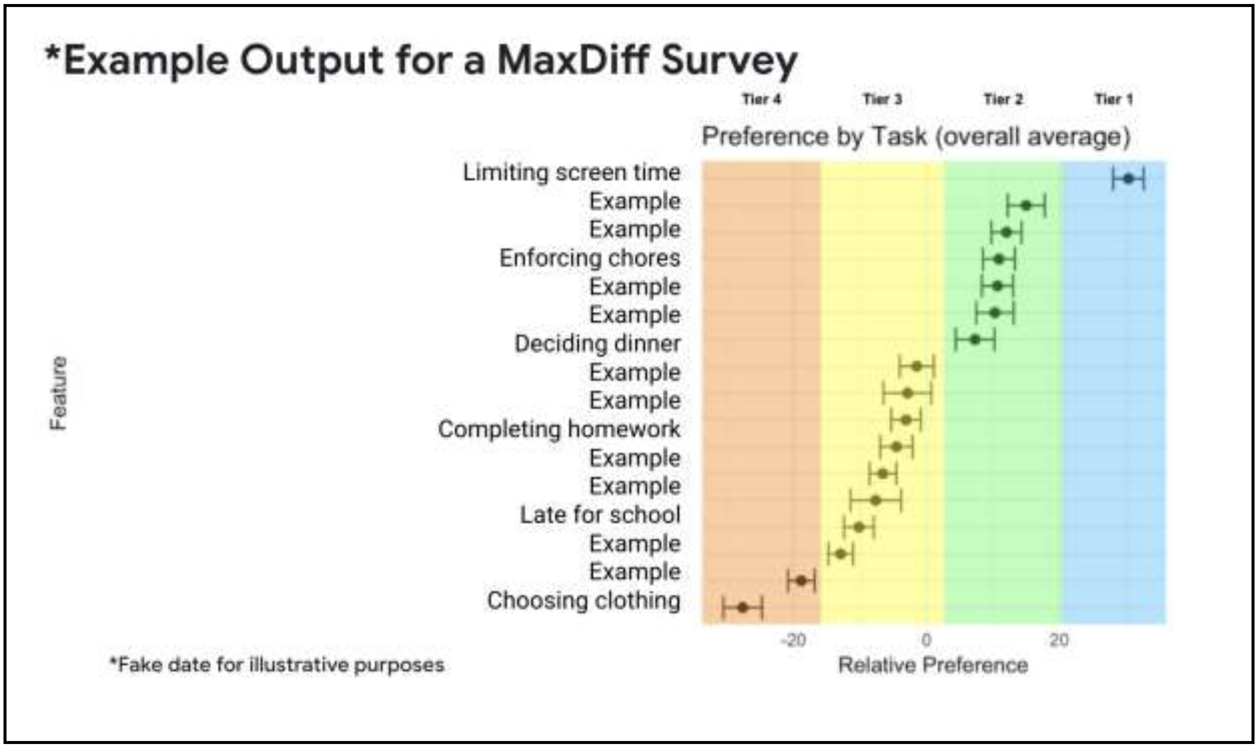
1. My kids could vote on their favorite meals at the start of the week.
2. They come shopping with us

The biggest hurdle to my solutions being adopted would be Getting the kids off Roblox

How important is it to solve the problem? 5/5

In Figure 7, an example output is shown for a MaxDiff survey based on common parenting challenges across a sample of parents. On the right-hand side is the most favorable problem to solve and on the left-hand side is the least preferred solution to solve.

Figure 7: Example Results for a MaxDiff Survey Based on CUJ Mad Libs



Discussion of Results

From looking at Figure 6 (the completed template) and Figure 7 (the MaxDiff graph), three observations become apparent. Firstly, the purpose of the template is not just to coax participants to fill in missing fields, but to open up a new line of questioning. For example, in Figure 6 we learn that there is a hurdle associated with the solution they are proposing—the children playing the game Roblox. Of course, the Product team doesn't have to take the solution offered by the participant by any means, but it does provide insight into another barrier that others may face.

The second takeaway from the template approach is that the template could either be empty or partially filled with a problem that the team wants to get more details on. To illustrate, perhaps the product team already knows that getting kids to complete homework is a chore; in this case this task can be prefilled such that participants can provide more context on this issue.

The third takeaway is that the importance of solving the problem (shown in Figure 6), provides the Researcher with a pulse on how iterations to the problem statement can affect the final MaxDiff ranking. A particularly satisfying feeling is seeing how a task that was initially rated as unimportant, can ultimately be rated more highly during the MaxDiff ranking because the User requirement is now better defined.

Conclusion

In summarizing the information presented in this paper, three reasons stand out for adding CUJ Mad Libs to your processes:

1. **Increases confidence in your scenarios:** It empowers the user to ground feedback in scenarios rather than slip into the mindset of performing a usability evaluation on wireframes. The advantage of this is that you may be more likely to gather nuances about how scenarios differ from one participant to another ahead of a MaxDiff survey.
2. **Easy to execute**: Since the technique does not require design assets it can be executed earlier in the process or appended as a section of another study. It also enables you to get feedback on existing problem statements that your team may have without closing the door on new problem statements that participants may freely offer on their own.
3. **It makes subsequent research steps easier:** When you pair CUJ Mad Libs with a MaxDiff survey, you can recruit better participants for the UI evaluation such as customers who thought the problem was important to solve or those who felt a problem was not important to solve because they already explored competitive solutions.

Finally, in this talk we presented one version of CUJ Mad Libs, but there are other variants that could be adapted to your needs; for example: have participants-prefill the templates with co-workers prior to the interview or explore gathering feedback via an unmoderated version of the technique.

Acknowledgements

We give special thanks to our Google XFN Partners who contributed to this study: Jolyn Yao (Product Manager) and Lauren Beckwith (Interaction Designer).

Jon Gass

Aideen Stronge

UPGRADE YOUR BRAND TRACKER USING THE POWER OF CONJOINT ANALYSIS

ALEXANDRA CHIRILOV
JAMES PITCHER
GFK

ABSTRACT

Most validation studies that compare conjoint preference shares with actual sales only focus on one single point in time. We go one step further and compare conjoint preference shares with real market shares over time. We test the ability of a simple brand-price conjoint to capture monthly fluctuations and the long-term trend in brand sales across three product categories (washing machines, laptops, and TVs) in two countries (Germany and UK) over 16 monthly waves of data. We find that neither conjoint analysis nor stated brand preference can accurately predict short-term fluctuations in market share due to external market factors that cannot be captured in a survey, having a large influence on changes in market performance. However, we demonstrate 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. Our findings suggest that conjoint analysis offers considerable benefits over traditional approaches to brand measurement and hence can enhance the decisions marketers make when investing in their brands.

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.

There is already wide evidence in the literature (Orme et al. (1999), Allenby et al. (2005), Hardt et al. (2017)), that conjoint shares can be closely aligned with actual sales. However, most validation studies only focus on one single point in time. In order to demonstrate that conjoint analysis is a suitable method for measuring Brand Equity in a tracking study, we need to estimate and validate conjoint preferences against market shares on a regular basis over time. To our knowledge, this has not been done before. Therefore, we conducted our own study to compare conjoint preference shares with real market shares (source: GfK POS Panel) on a monthly basis in a brand tracking survey to answer five main questions:

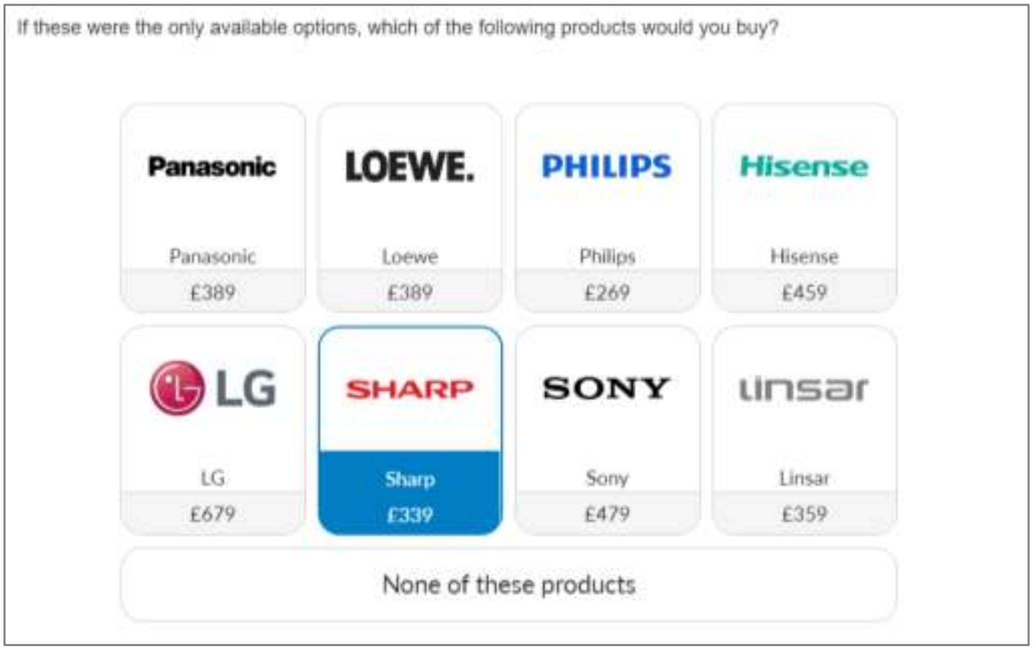
1. At any point in time, how close are conjoint preferences to real market shares?
2. How stable are conjoint preferences over time?
3. How well can conjoint preferences capture long-term changes in market shares over time?

4. How well can conjoint preferences capture the monthly fluctuations in market shares?
5. How do conjoint preferences compare to traditional stated brand preferences on all of the above?

Research Design

We fielded six CBC (Choice-Based Conjoint) surveys with online respondents in Germany and the UK in three distinct product categories: TVs, laptops, and washing machines. The study has been running on a monthly basis since September 2019, with 200 respondents per cell per wave. Our analysis is based on 16 months' worth of data.

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:



If these were the only available options, which of the following products would you buy?

Panasonic £389

Loewe £389

Philips £269

Hisense £459

LG £679

Sharp £339

Sony £479

Linsar £359

None of these products

Paper Tissues study

Our results also refer to a study of 13 brands of paper tissues 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 had 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.

Scaling of Shares

Conjoint shares of preference were calculated without the "none" option included, so they summed to 100%. Stated preference and market shares were rescaled to sum to 100% across all brands tested in the conjoint.

Share Adjustments by Distribution and Awareness

As well as comparing the raw conjoint and stated preferences with market shares, we also adjusted them by distribution, and in the case of conjoint, separately by awareness. Shares were adjusted by distribution by multiplying the aggregated conjoint and stated preferences by the total weighted distribution of the brand (expressed as a percentage) and then rescaling the shares to sum to 100%. Conjoint preferences were independently adjusted by awareness in the same way. Note that stated preferences are already inherently adjusted by awareness as the traditional brand funnel assumes respondents can only prefer a brand that they are aware of.

RESULTS

Cross-Sectional Validation

In our cross-sectional validation we compare, at any point in time, how close the shares of each brand we obtain from our conjoint models and stated preference are to the real-world market shares of those brands.

Figure 1 shows a comparison of the Mean Absolute Errors (MAEs) for both conjoint preferences and stated preferences with market shares, for all brands within all 6 cells, across all 16 waves. The MAEs for conjoint preferences are notably lower than stated preferences for the UK cells and for Washing Machines in Germany. In Laptops and TVs in Germany, there is little difference in the MAEs of the two approaches.

Figure 1: Mean Absolute Errors for Conjoint Preferences and Stated Preferences with Market Shares.

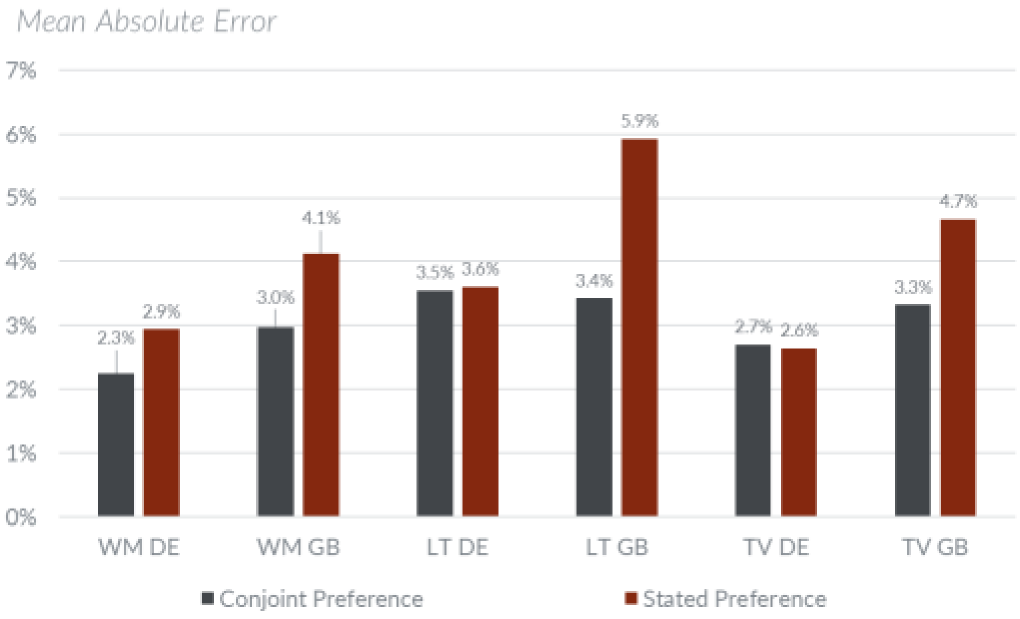


Figure 2 shows a plot of the conjoint preference shares against market shares, for all brands within all 6 cells, across all 16 waves. There is a very strong relationship between the two, with a very high correlation of 85%. However, we see that conjoint preferences consistently under-estimate the market share of HP laptops.

Figure 2: Correlation between Conjoint Preferences and Market Shares.

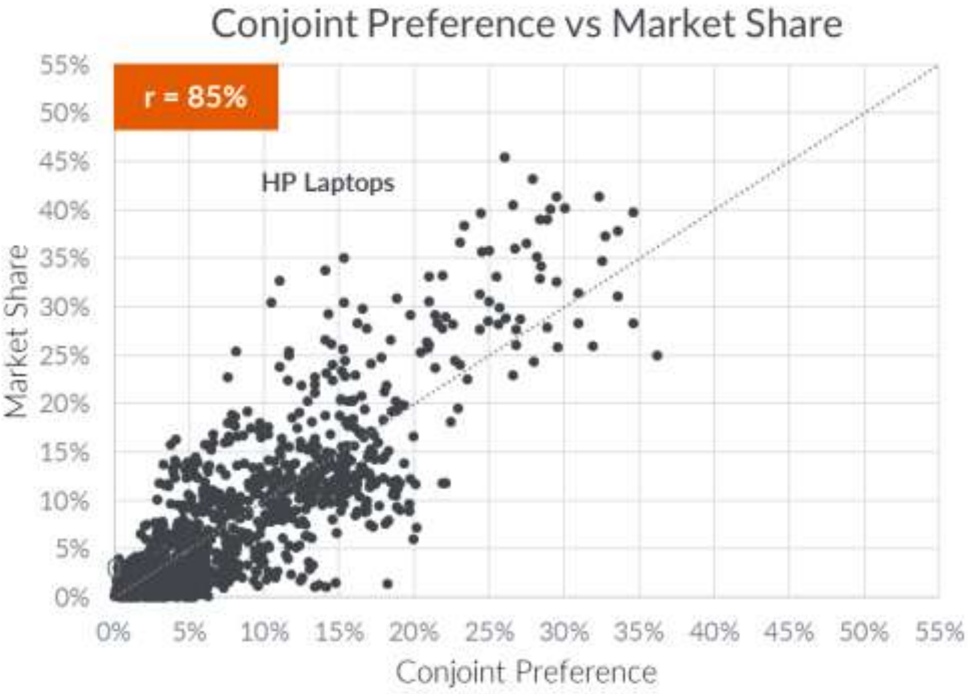


Figure 3 shows the same plot but for stated preference versus market shares. With a correlation of 78%, the overall relationship is strong, but slightly weaker than conjoint preference versus market shares. We also observe large differences in shares for some brands. Like conjoint preference, stated preference underestimates the market share of HP laptops. But we also see that stated preference overestimates the shares of big brands and premium brands, such as Apple, Samsung, and Miele.

Figure 3: Correlation between Stated Preferences and Market Shares.

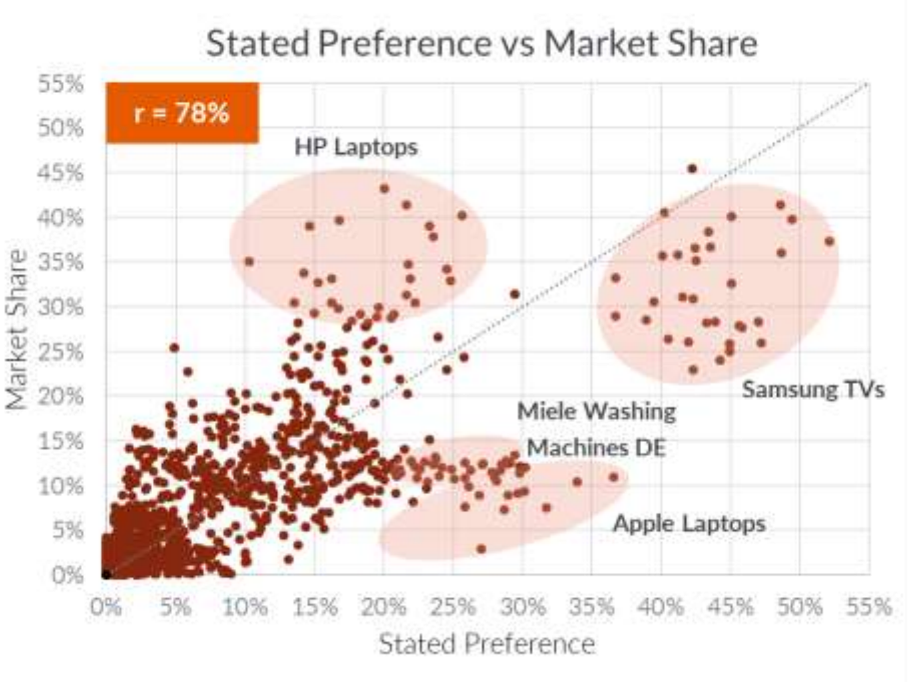


Figure 4 demonstrates how stated preference (21.0%) can vastly over-estimate the market share (12.9%) of large brands such as Apple (laptops) that are successful in many different product categories. In contrast, conjoint preferences (13.1%) provide more realistic estimates of the share of these big brands.

Similarly, Figure 5 demonstrates how stated preference (22.1%) can vastly overestimate the market share (12.3%) of very premium brands such as Miele (washing machines) that have a high price tag compared with the rest of the category (Figure 6 shows the market average prices of each brand of washing machine). In contrast, conjoint preferences (11.6%) provide more realistic estimates of the share of these premium brands.

Figure 4: Comparison of Conjoint Preference, Stated Preference and Market Share of Apple Laptops in Germany (*September 2019, Wave 1).

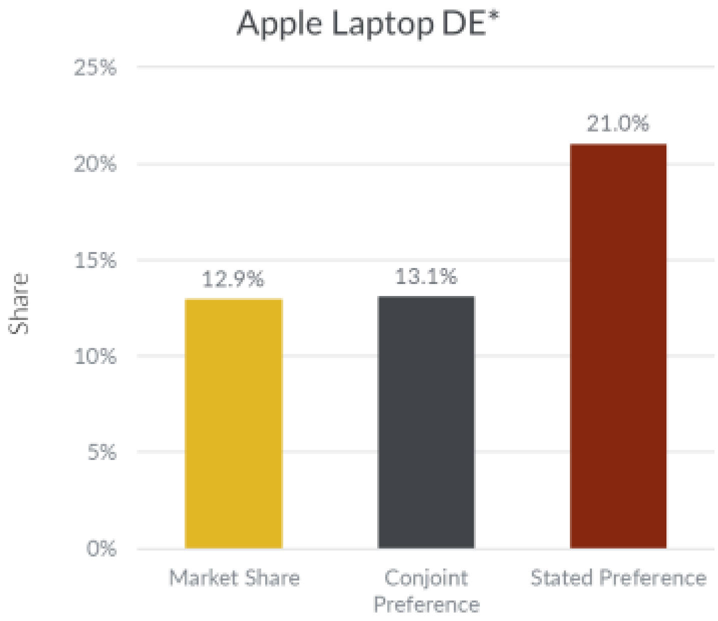


Figure 5: Comparison of Conjoint Preference, Stated Preference and Market Share of Miele Washing Machines in Germany (*September 2019, Wave 1).

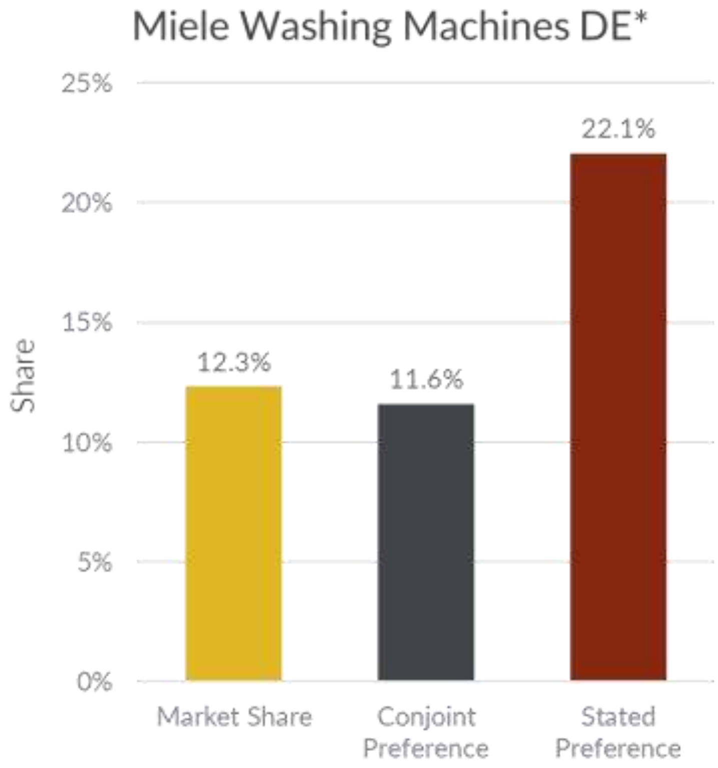


Figure 6: Comparison of Prices Tested for Brands of Washing Machines in Germany.

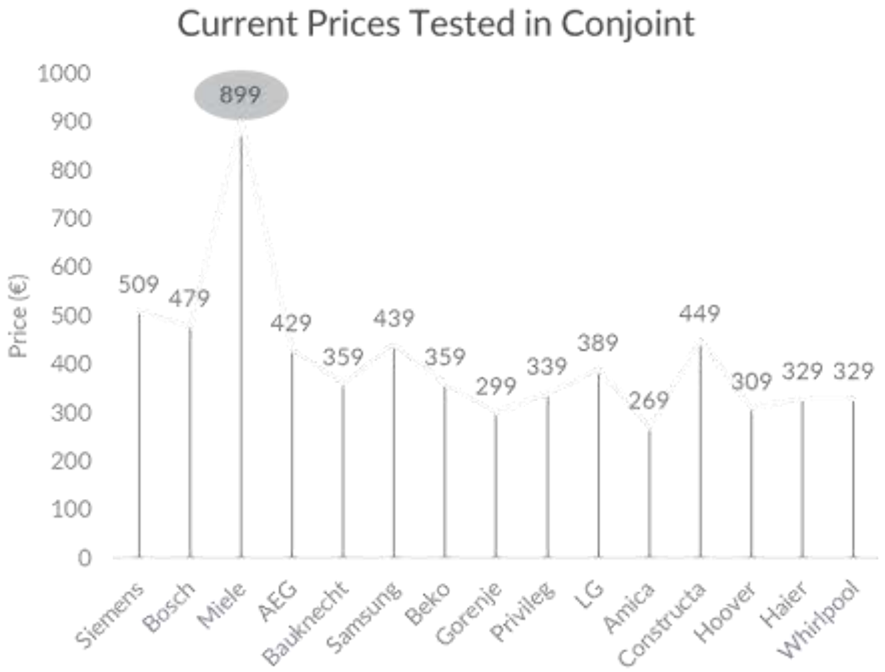


Figure 7 shows the MAE for conjoint preferences is reduced when adjusted by distribution, most notably for laptops in Germany, which has the highest error. However, when stated preferences are adjusted by distribution, the impact in lowering errors is reduced. In fact, the error increases for 3 out of the 6 cells.

Figure 8 shows the MAE for conjoint preferences is not improved when adjusted by awareness. Furthermore, Figure 9 shows that, in a separate study of 13 paper tissue brands in Germany, while adjusting conjoint preferences by distribution reduces the MAE (from 2.5% to 2.1%), adjusting conjoint preferences by awareness increases the MAE (3.2%).

Figure 7: Mean Absolute Errors for Conjoint Preferences and Stated Preferences with Market Shares, Adjusted by Distribution.

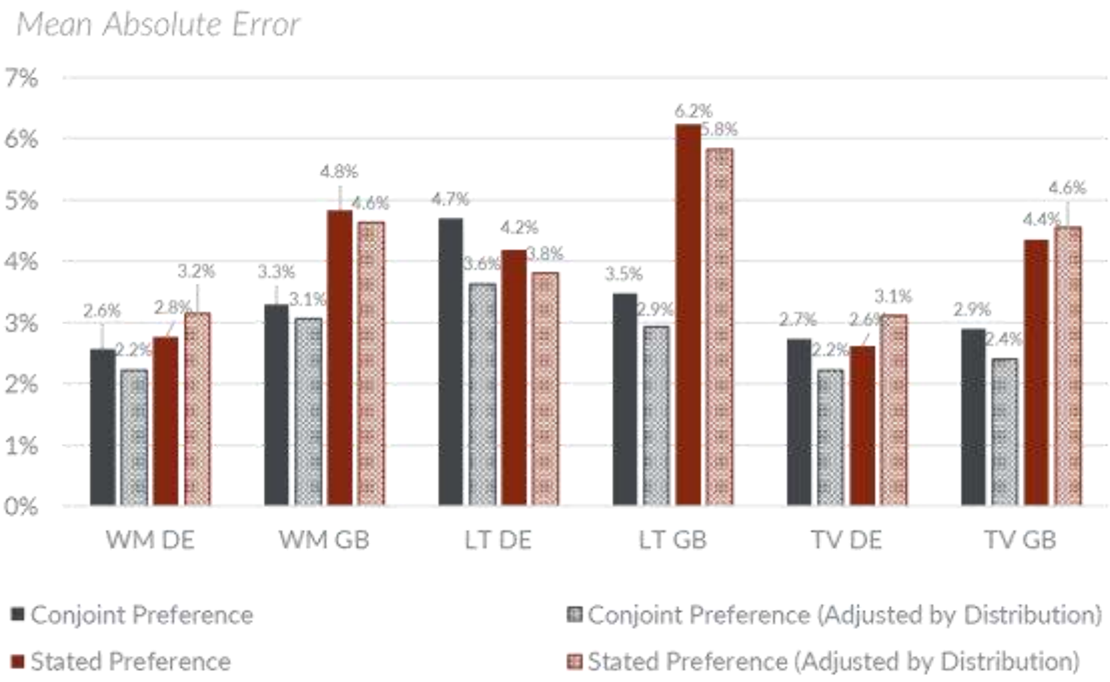


Figure 8: Mean Absolute Errors for Conjoint Preferences with Market Shares, Adjusted by Awareness.

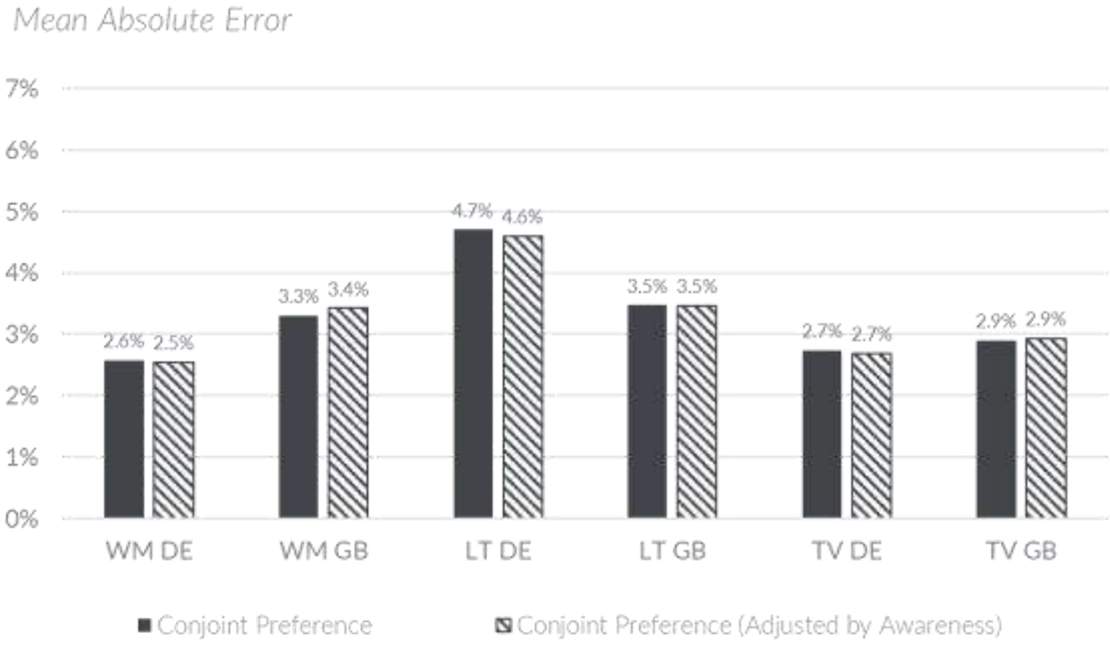
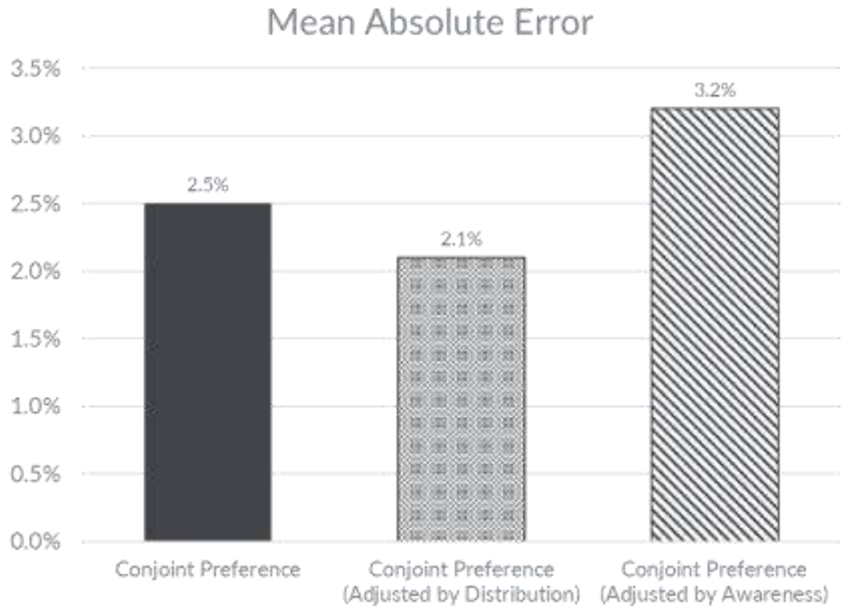


Figure 9: Study of 13 Brands of Paper Tissues in Germany: Mean Absolute Errors for Conjoint Preferences with Market Shares, Adjusted by Distribution and Awareness.

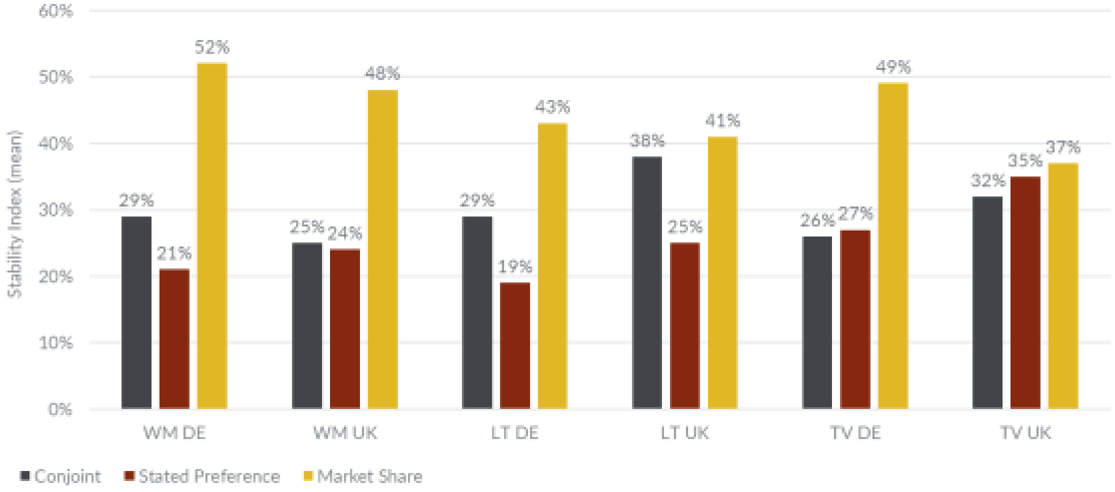


Longitudinal Validation

In our longitudinal validation, we first compare the shares of each brand we obtain from our conjoint models and stated preference in terms of how stable they are over time and how their stability compares with that of real-world market shares.

To assess stability, we calculated a “Stability Index” that represents the percentage of the relative monthly changes in share that fall within +/-10%. Figure 10 shows a comparison of the average Stability Index for conjoint preferences, stated preferences, and market shares across the 6 cells and 16 waves. The Stability Index is highest for the market shares in all cells and is notably much higher than both survey metrics in most cells. In laptops, and washing machines in Germany, conjoint preferences have a higher Stability Index than stated preference but there is little difference between the two methods in the other three cells.

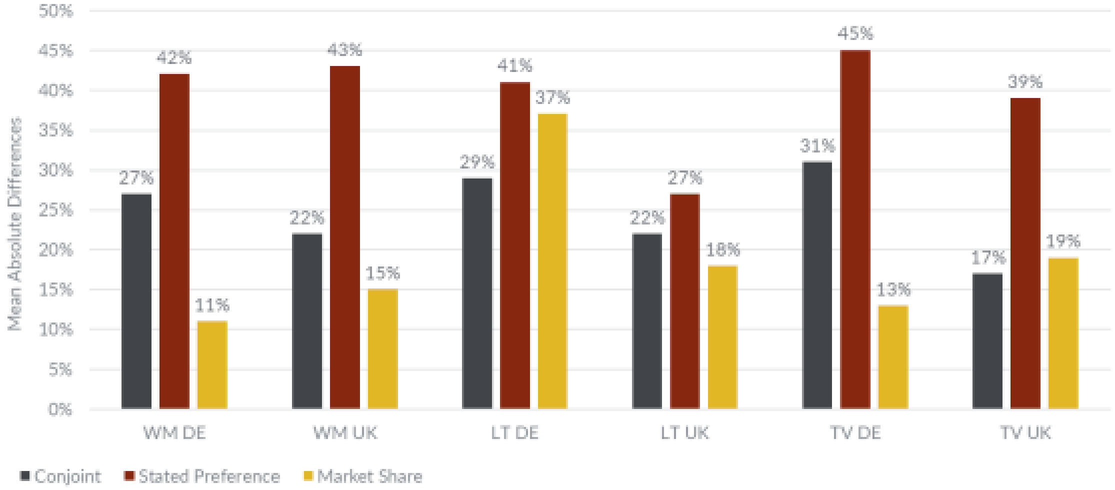
Figure 10: Average Stability Index for Conjoint Preferences, Stated Preferences, and Market Shares.



N=53 brands, brands with monthly values above 0.01%.

Figure 11 shows the average absolute relative change in conjoint preferences from month to month are, in most cells, much lower than that of stated preference and closest to the low average relative changes we observe in market shares. These figures, along with the Stability Index, suggest that conjoint preferences are generally more stable than stated preferences, and market shares are the most stable of all.

Figure 11: Average Absolute Relative Change in Monthly Conjoint Preferences, Stated Preferences and Market Shares.



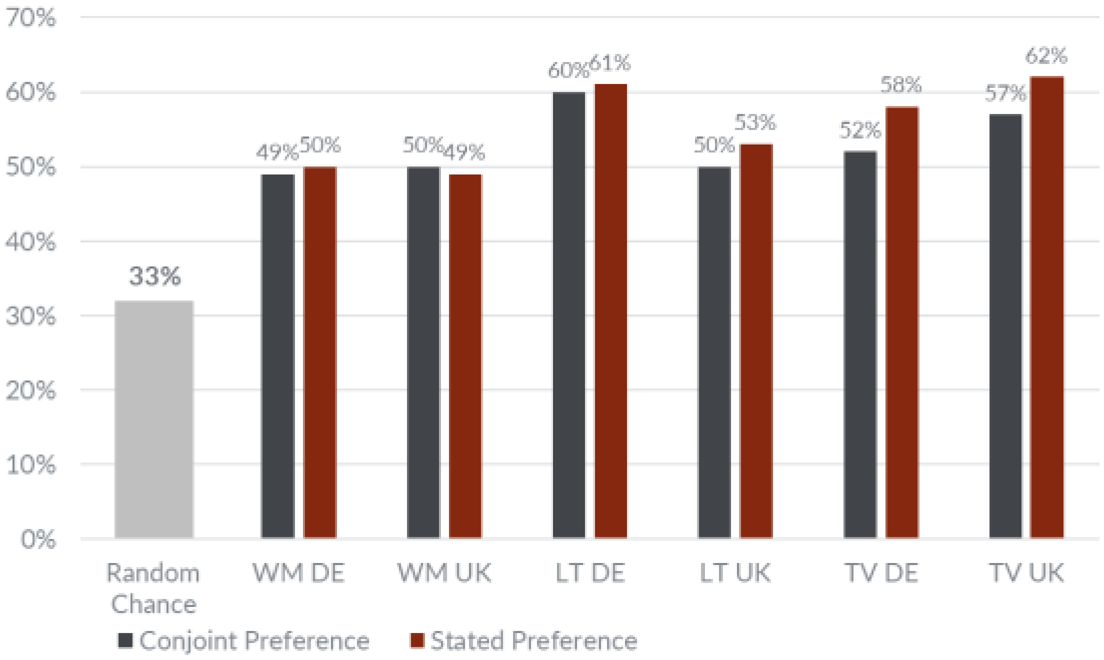
N=53 brands, brands with monthly values above 0.01%.

Next, we first compare the extent to which the month-to-month changes in the shares of each brand we obtain from our conjoint models and stated preference match the monthly changes observed in real-world market shares.

We computed the monthly hit rates for conjoint preferences and stated preferences with respect to their ability to match the directional change in market share of brands in all 6 cells over the 16-month time period. If the conjoint preference and market share for a brand both increase or decrease in a given month compared to the previous month, then that represents a "Hit." If one increases and the other decreases, then that is a "Miss." The hit rate is the percentage of "hits" we observe.

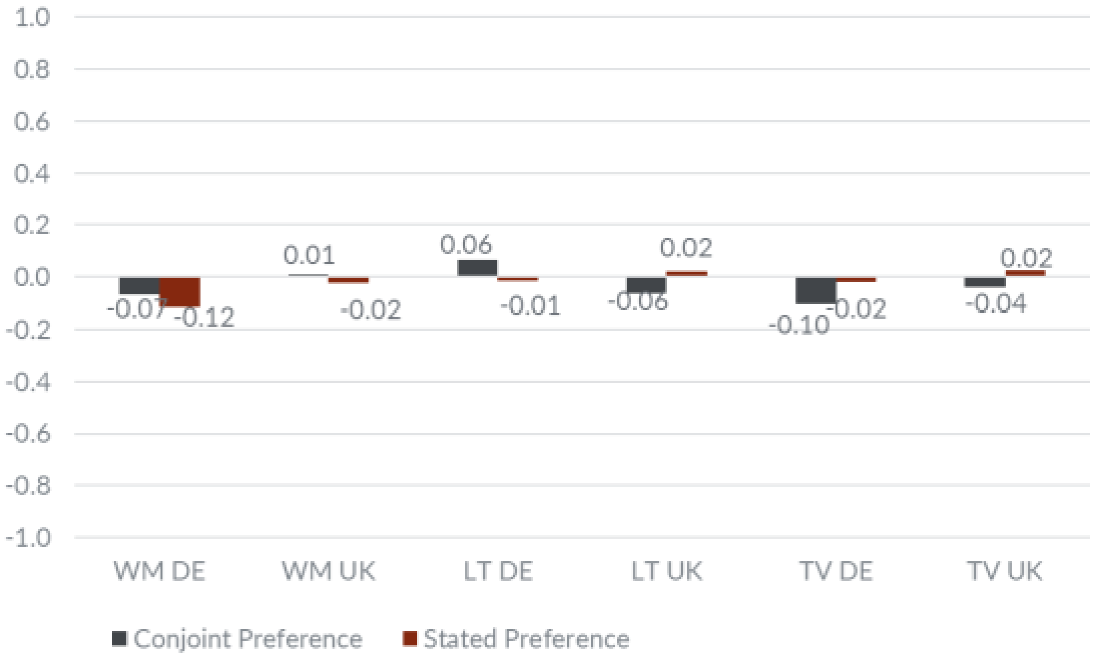
Figure 12 shows there is little difference between the hit rates for conjoint preferences and stated preferences in all cells. Both methods only correctly match 50%-60% of directional changes in market share. Although this is better than random chance (33%), only 51% of brands have an average hit rate of greater than or equal to 50%. Furthermore, Figure 13 shows there is almost no correlation between monthly changes in conjoint preferences and stated preferences and monthly changes in market shares.

Figure 12: Monthly Hit Rates: Percentage of Times the Directional Change in Conjoint Preferences and Stated Preferences Match the Changes in Market Shares.



N=53 brands, brands with monthly values above 0.01%.

Figure 13: Kendall Correlations of Monthly Change in Conjoint Preferences and Stated Preferences with Monthly Change in Market Shares.

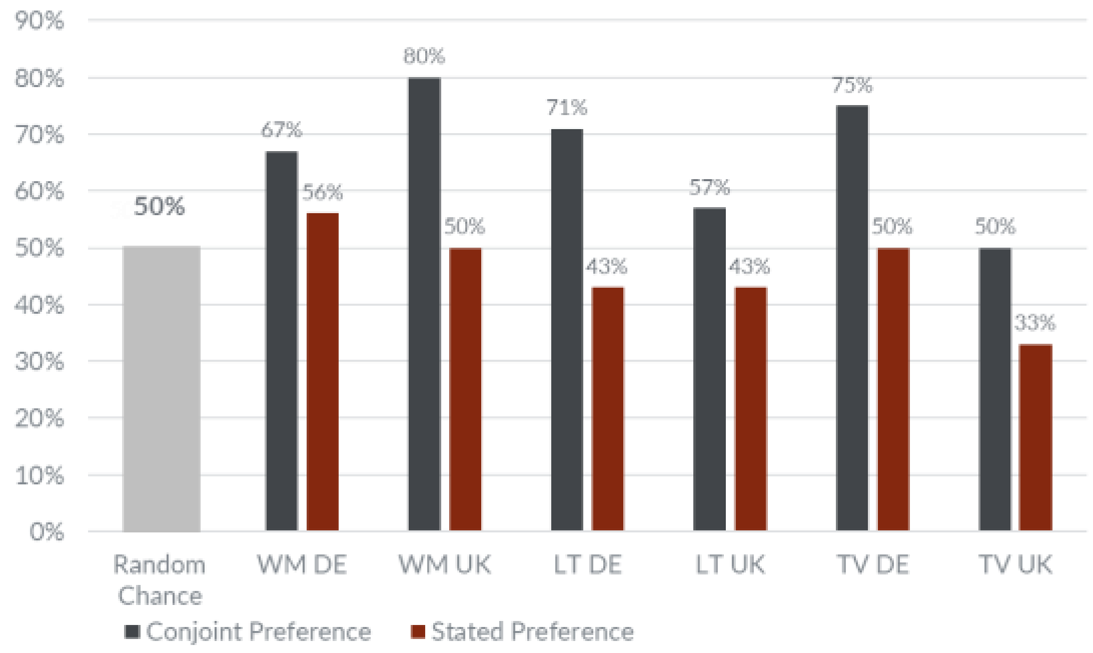


N=53 brands, brands with monthly values above 0.01%.

Finally, we compare the ability of conjoint preferences and stated preferences to capture long-term changes in market shares over time. We do this by correlating the trends in conjoint preferences and stated preferences with the trends in market shares for each brand, across the 16 months. For example, if the trend in a brand's conjoint preference increases over time along with the trend in its market share, then the trends are said to correlate. Similarly, the trends in conjoint preference and market share correlate if both decrease over time. However, if one increases and the other decreases, then they are said to not correlate.

Trends in conjoint preferences and market shares correlate amongst 68% of all brands in all 6 cells. This represents a 45% increase in the ability to predict long-term trends in market share compared with stated preference, where only 47% of brand trends correlate with the trends in market share. Figure 14 shows that correlations in trends between conjoint preferences and market shares are notably higher in all cells compared with the correlation in trends between stated preferences and market shares.

Figure 14: Percentage of Brands for Which the Trend in Conjoint Preferences or Stated Preferences Correlates with the Trend in Market Shares.



N=49 brands, brands with a market share <1.5% were removed because the results were highly unstable.

Discussion

Conjoint preferences reflect market shares better than stated preferences.

We have observed that conjoint preference is a better predictor of a brand's market share than stated preference. At any point in time, there are fewer differences between conjoint preferences and market shares, compared with stated preferences. Furthermore, conjoint preferences are better at capturing the long-term trend in a brand's market share. If we observe shifts over time in a brand's conjoint preference, it is likely we will also observe similar shifts in that brand's market share. Stated preference is a less reliable predictor of such changes.

One reason why conjoint preferences are closer to the sales reality is that the conjoint exercise mimics how consumers choose brands in the real world. Whereas stated metrics are a poor reflection of that reality. Simply listing a series of brand names on a screen and asking the respondent to select which brand they prefer is far from the reality of how that respondent would actually go about choosing a brand in the real world. In contrast, the conjoint exercise places respondents in a realistic buying-like scenario. This added realism leads to better data on the choices that consumers make in their everyday lives.

Stated preference greatly overestimates the market shares of large brands such as Apple and Samsung, that are successful in many different product categories. Conjoint analysis provides far more realistic estimates of the share of these big brands. The likely reason for this is that conjoint does a much better job than stated preference of anchoring consumer choices within the category of interest. For example, when respondents evaluate Apple within the laptop market, their stated preferences are influenced by Apple being such a strong brand in general. They are thinking about the Apple brand in more general terms and not evaluating Apple specifically within the context of buying a laptop. Since Apple enjoys very strong market success in other categories, such as mobile phones and tablets, this creates a halo effect that leads to the preference for Apple to be overstated. The conjoint exercise reduces this halo effect by placing respondents in a realistic buying-like scenario. Hence, brands are evaluated specifically within the category of interest. Respondents evaluate Apple laptops, not Apple in general. We, therefore, obtain much more accurate shares for large brands that play in multiple categories, such as Apple, using conjoint analysis compared with stated preference.

Stated preference also greatly overestimates the market shares of very premium brands, such as Miele (washing machines), that have a high price tag compared with the rest of the category. Again, conjoint analysis provides far more realistic estimates of the share of these expensive brands. This is because the conjoint exercise shows brands at their associated prices which adds realism. It reflects the fact that Miele is much more expensive than any other brands in the market and hence respondents take this into account when they make their choice of brand. In contrast, stated preference does not show brands with any associated prices. Hence, respondents are more likely to choose Miele because there is no additional cost associated with the Miele brand. Conjoint reflects the fact that many of the respondents who state that they prefer Miele would never pay that much for a washing machine in real life and therefore provides preferences for premium brands that are much closer to real-world market shares than stated preferences.

Conjoint preferences will never be the same as market shares.

Although conjoint preferences come close to market shares, they will never exactly match. There will always be differences because market shares are subject to external market factors that are not captured within the conjoint exercise. For example, product placement in store, space taken up on the shelf, recommendation of sales staff, and in-store promotional material such as display signs, can all influence the choice of brand at the moment of purchase.

Another important factor that influences market performance is physical distribution. We generally see that brands which are available to be bought in more places enjoy higher sales. However, conjoint assumes all products are available to all consumers equally.

Of course, it isn't just conjoint analysis that is unable to capture these external market factors. They are a feature of any survey question. The reason why both conjoint preference and stated preference underestimate the market share of HP laptops is likely to be because both fail to capture external factors that inflate HP's market share. For example, HP has the highest distribution in the market and there is some evidence that HP laptops are recommended by sales staff more than other brands.

These external market factors also explain why it is difficult to align survey data, be it conjoint preferences or simple stated preferences, with monthly changes in market share. For example, a brand's market share might increase in a particular month because of a short-term advertising campaign or price promotion that the brand ran to activate sales. Therefore, it is perhaps unreasonable to expect conjoint preferences alone to ever be able to accurately predict monthly fluctuations in market shares.

However, differences in conjoint preference and market shares can provide insights in and of themselves. For example, if conjoint preference predicts a higher share for a brand than what we see in market, we can draw the conclusion that there are some in-market factors that are preventing the brand from reaching its full potential. Conversely, if conjoint preference predicts a lower share for a brand than what we see in market, the brand may be at risk of losing share in the future if market conditions change.

Conjoint preferences adjusted for distribution are even closer to market shares.

As just discussed, external market factors that are not captured within the conjoint exercise influence market shares, and the level of distribution of a brand or product is often a large indicator of its share in the marketplace. Therefore, it is reasonable to assume that adjusting the conjoint preferences for brands by their different levels of physical distribution would result in shares that more accurately reflect market shares. Hence why it is a common method of adjustment used by practitioners.

In our study, we indeed see that adjusting by distribution brings conjoint preferences even closer to the brand market shares. However, we do not see the same decrease in error for stated preference, when adjusted by distribution. In some cases, the error increases. This is because brands with the highest distribution are the same large brands for which the market shares are already overestimated by stated preference. Therefore, adjusting for distribution exacerbates this overprediction of such large brands.

Consumers sometimes buy brands they've never heard of.

Another common way to calibrate conjoint preferences is using awareness. However, this does not bring conjoint preferences closer to market shares. The MAEs did not improve for any cell. This is because the conjoint exercise already captures the role of awareness of brands in respondent choices. If respondents are not aware of a brand, they can simply not choose that brand in the conjoint exercise.

Conversely, respondents have the option to choose a brand they are not aware of. This is important as consumers sometimes buy products from brands unknown to them. If a product looks acceptable and is offered at a fair price, a person may buy a product from a brand they have never heard of. This is especially true in low involvement categories, where prices don't represent a significance financial investment. Hence, consumers are willing to risk buying a brand unknown to them. This is why in our study of 13 brands of paper tissues in Germany, we found that calibrating the conjoint preferences by awareness made the MAE versus market shares worse, not better. It is therefore wrong to always assume that consumers do not buy brands they are not aware of, as is the case with traditional brand funnels.

Conjoint preferences have better stability than stated preferences.

Our results show that the average relative change in stated preference from month-to-month is over double that of market shares. This lack of stability makes it an unreliable measure of the current market share of a brand. However, conjoint preferences are much more stable, and the monthly fluctuations are much closer to the changes we see in real market shares, making it a much more reliable measure of consumer preferences. Note that this stability was achieved using a minimum sample of just 200 respondents per month. It is expected that stability of our conjoint preferences would be even closer to the stability of market shares if larger sample sizes are used.

Implications for Brand Tracking

Marketers expect their brand metrics to be linked to sales.

An organisation's brand is among its most valuable assets. Building a strong brand is key to gaining a good market position, brand premium, and sustainable growth. However, marketers often struggle to get investment in their brand building activities because they are expected to provide hard numbers to measure the commercial impact of such brand activities. In this paper we have shown that conjoint preferences are more closely linked to sales than traditional stated metrics, and hence can provide a new, superior way for marketers to assess the impact of their brand activities on sales.

Conjoint preferences align with modern brand theory on how we choose brands.

Conjoint preferences reflect the fact that brand buying is probabilistic. In his well-known book, "How Brands Grow," Professor Byron Sharp states:

> *"…individual brand memories, like our brand buying, are probabilistic. We each have a steady, ongoing propensity to think something, and for most of our beliefs that propensity is not 100%."*

Sharp is essentially saying that we don't think about brands in absolute terms. We don't have an absolute preference for one brand over another. Instead, we have a *propensity* to prefer one brand over another. Whereas, stated preference unrealistically asks respondents to state, with 100% probability, which brand they prefer; it is this *propensity* to prefer a brand which is measured by conjoint preferences. Conjoint preferences represent the *extent* to which consumers prefer each brand.

Conjoint preferences provide richer data.

A major benefit of conjoint preferences measuring the *extent* to which consumers prefer each brand is that, compared with stated preference, conjoint preferences provide far more granular data. Stated brand preference only provides information on which brand a respondent prefers. It doesn't tell us the extent to which a respondent prefers their chosen brand or their relative preference for all other brands. In contrast, conjoint preferences are a scaled measure of preference that is comparable across all brands for each individual respondent. This enhanced richness of conjoint preferences provides great benefits when running additional diagnostic analyses, such as running Key Drivers Analysis and segmenting respondents, to generate further insights for brands.

Conjoint analysis provides many more useful ways to diagnose a brand.

Conjoint analysis provides many useful additional brand insights compared with traditional stated brand metrics. For example, we can learn how consumers switch between different brands; how likely is a brand to gain/lose share from/to its competition? From/to which competing brands is a brand most likely to gain/lose share? We can also calculate brand elasticities from conjoint data, gaining insights into the extent to which brands can charge a premium. This is important and something that is often overlooked by traditional brand measures, as growing market share is not the only way to increase profitability. Building brand premium can have an even bigger impact on a brand's profitability.

Conclusion

Conjoint analysis is traditionally deployed on an ad hoc basis and with good reason. It is a highly complex method that is difficult to apply in a brand tracker due to its high cost and high level of customisation. However, due to advances in technology, we have shown how it is possible to automate a simple brand-price conjoint exercise so it may be used on a continuous basis to track the performance of brands. We have shown that using conjoint analysis in this way offers considerable benefits over traditional approaches to brand measurement, and hence can enhance the decisions marketers make when investing in their brands.

Alexandra Chirilov

James Pitcher

References

Orme et al. (1999) Predicting Actual Sales with CBC: How Capturing Heterogeneity Improves Results

Allenby et al. (2005) Adjusting choice models to better predict market behaviour

Hardt et al. (2017) Reconciling Stated and Revealed Preferences

Professor Byron Sharp (2010) How Brands Grow

RESPONDENT QUALITY: IDENTIFYING BAD RESPONDENTS FROM MAXDIFF RESPONSE PATTERNS

JANE TANG
MONA FOSS
ROSANNA MAU
BOOTSTRAP ANALYTICS

INTRODUCTION

Although much work has already been done in the effort to identify poor quality respondents in market research data—be they unengaged respondents or nefarious bots—questions remain about which methodological approach is most effective and what is the best way to implement it. Our work looks to build on that of past researchers, test the various approaches and evaluate their effects using real-world data.

For the purposes of this research, a "bad" respondent is one who provides random response patterns. To identify these random responders, we can look for patterns in Automated Test Respondent (ATR) data using a MaxDiff exercise and identify respondents in the real data with similar patterns. In Hierarchical Bayes models, ATR data has a lower RLH (root likelihood) fit score than real data, making it useful for devising a cutoff point. Past researchers have also used Latent Class MNL (multinomial logit) models that included ATR data. The latent classes that included the majority of the ATR cases would be deemed to have suspect response patterns and true respondents falling into those classes would be flagged for removal. More recently a Scale-Adjusted Latent Class (SALC) model has been used without the need for ATR data. Respondents in the class constrained to have scale equal to zero are identified as random behaving.

By removing these respondents, our goal is to produce research results with less noise and stronger signal.

PAST RESEARCH

Hoogerbrugge and de Jong (2019) introduced the idea of using ATR data to help identify true respondents who provide poor quality choice data using a combination of a root likelihood (RLH) score and Latent Class MNL modeling. The authors used the RLH scores and latent class probabilities in a logistic regression to predict random versus real respondents. Although their LC model clearly identified classes that were dominated by the ATR data, they found the logistic regression modeling helped to better identify the real from the random respondents.

The authors also spent some time determining a threshold below which respondents would be removed from the data. They settled on an equal misidentification rule to identify the cutoff, which maintained a balance between "throwing away real respondents" and "keeping bad respondents."

Additional work from Orme (2019) recommended a 95th percentile of RLH cutoff based on an HB model using only ATR data. Any random responder completing the survey would have a 95% likelihood of falling below this cutoff level. He also suggested incorporating other survey-level data, such as time to completion and straight-lining, to flag other candidates for deletion. Orme cautioned that sparse MaxDiff data will regularly produce high RLH scores, making it more difficult to distinguish between random and conscientious responders.

Finally, Chrzan (2020) used an artificial dataset of respondents programmed to be "real" or "random." He tested both the HB-RLH and the LC-MNL approaches, along with a new scale-adjusted latent class (SALC) model. SALC models identify differences in scale, based on response error reflected in the logit scale parameter, rather than just differences in preference. They can be used to identify respondents whose utilities imply random choosing. Chrzan's evaluation focused on classification—the rate of true positives (the ability to identify random responders) to the rate of false positives (misidentifying real responders as random)—to assess the three models under varying scenarios, from robust to sparse data. He concluded that the RLH method dominated both LC-MNL methods.

Gaps in the Research

These earlier papers focused on minimizing misidentifications—real respondents being identified as random and ATR respondents being identified as real. However, minimizing misidentification is not the ultimate goal of the exercise. Our research seeks to move beyond misidentification and look at the impact of removing bad data on the research results. If the goal is to improve our data by removing the noise, we may be better served by removing *all* of the real respondents who behave like random, even if we end up misidentifying some real data.

Methods

Our study investigated 3 methods:

1. HB-RLH with 95th percentile cutoff (going forward, we will label this **RLH only**);
2. Logistic regression to predict the probability of being ATR using RLH score and LC probabilities (we will label this **RLH+LC**); and
3. Logistic regression to predict the probability of being ATR using RLH score and scale class probability (we will label this **RLH+SALC**).

For both of the logistic regression models, we further investigate if we should use:

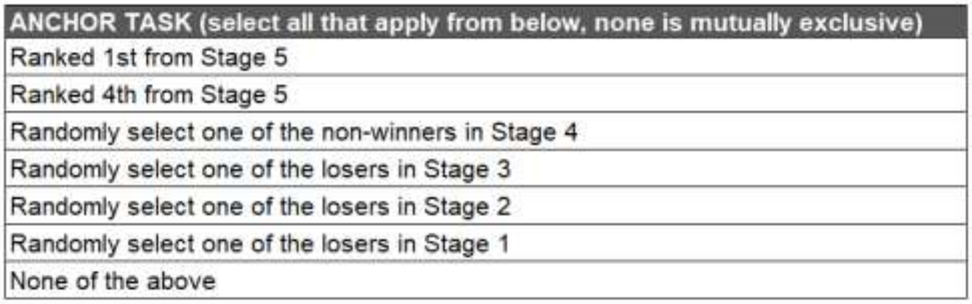
a. The 95th percentile cutoff based on predicted probability from the regression (we will label this **95% cutoff**); or
b. The equal misidentification rule, used by Hoogerbrugge and de Jong, to determine the cutoff (we will label this the **Equal MisID cutoff**).

We used three sets of real-world data to conduct this research:

- n=2,000 from Country A, which was generally believed to be of good quality;
- n=4,000 from Country B, which potentially included a lot of "bad" respondents;
- n=1,800 ATR cases.

Each real and ATR respondent completed an Anchored Tournament (aka Adaptive) MaxDiff with 24 items, using the following design:

Stage	# of Items	# of Tasks	# of Options	Task	Discard
1	24	8	3	Winner/Loser	Loser
2	16	4	4	Winner/Loser	Loser
3	12	4	3	Winner/Loser	Loser
4	8	4	2	Winner	Non-winner
5	4	1	4	Rank	



ANCHOR TASK (select all that apply from below, none is mutually exclusive)
Ranked 1st from Stage 5
Ranked 4th from Stage 5
Randomly select one of the non-winners in Stage 4
Randomly select one of the losers in Stage 3
Randomly select one of the losers in Stage 2
Randomly select one of the losers in Stage 1
None of the above

We chose to use the tournament MaxDiff because it is what we use most often in our day-to-day research. However, it has a downside. Since there is an internal built-in logic to tournament MaxDiff data, even the ATR data itself is not completely random. We were interested to see if any of the methods for identifying bad respondents would work in this case.

Respondents also completed a concept testing exercise, rating 3 of the 24 items on a variety of measures.

Research Objectives

The main objective of our research was to determine which of the three methods (RLH only, RLH+LC, RLH+SALC) and two cutoff approaches (95% or Equal Misidentification) is most effective in terms of reducing the noise in the data. To measure this noise reduction, we look at improvements in the differentiation in MaxDiff results, as well as improvements in the differentiation in concept test results after removing respondents flagged by the various methods.

Secondly, we sought to determine if it is better to run a separate HB model for the ATR data, or to combine it with the real data and run both together in one HB model. We also examine how large of an ATR dataset is necessary, by testing sample sizes between 15% and 90% of the real data.

Using the two country datasets described, we varied the amount of ATR data used to come up with the following runs:

- Country A, n=2000 + 300 ATR (15%)
- Country A, n=2000 + 1800 ATR (90%)

- Country B, n=2000 + 300 ATR (15%)
- Country B, n=4000 + 600 ATR (15%)
- Country B, n=4000 + 1800 ATR (45%)
- Country B, n=2000 + 1800 ATR (90%)

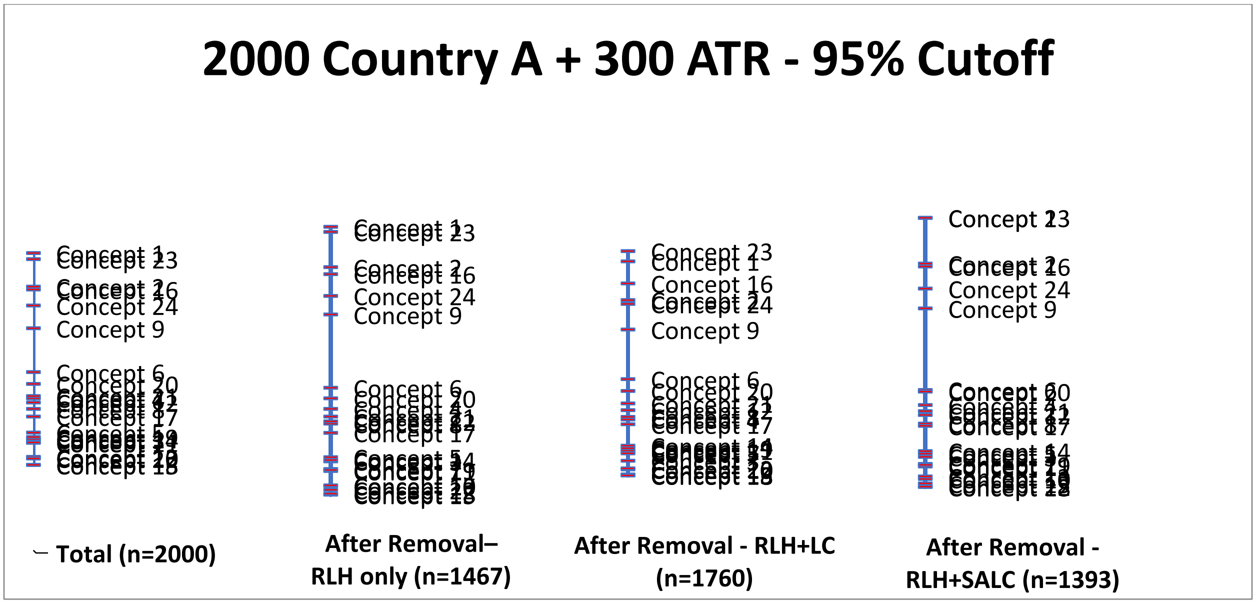
For each of these datasets (real and ATR data combined), we ran an HB model, a latent class MNL model with 20 classes, a SALC model with 20 classes and 2 scale classes, as well as logistic regressions for the LC and SALC models to predict ATR versus real data using RLH score and class or scale class probabilities. It should be noted that while the SALC model does not require ATR data, we did include it as we wanted to keep the datasets consistent in order to compare across methods.

Additionally, we did not attempt more than 2 scale classes in the SALC models due to the very long run time required. We used Sawtooth Software's CBC-HB for the HB runs and Statistical Innovations' Latent Gold Software for the LC and SALC models.

We also ran HB models separately for each of the real and ATR datasets described above. The ATR-only runs were used to determine the 95th percentile RLH cutoff for the real data runs.

Findings

To compare any improvements in differentiation from one method to another, we calculated the standard deviation across the 24 beta values (after logit transformation) for each sample and indexed this to the standard deviation in the total sample. The indexed value shows any improvement in differentiation (if greater than one) in the cleaned data. We use this index throughout these findings to compare the performance of the various methods, where a larger index score indicates better improvement.



St. Dev. Index to Total: 1.28 1.07 1.29

These side-by-side thermometer charts plot the MaxDiff scores for each removal method using Country A data with 15% ATR. The longer the vertical bar, the greater the differentiation in the betas. In this case, the RLH only and RLH+SALC methods result in very similar improvements to differentiation, as seen by the length of the thermometer charts, as well as the index scores. RLH+LC does less well.

RLH Only versus RLH+SALC

Looking more closely at these two methods in particular, and at all six data sets, we find that the fewer respondents retained after data cleaning, the better the differentiation in MaxDiff scores, summarized by the index scores. Both methods offer significant improvements in differentiation of MaxDiff beta scores.

	RLH only 95% cutoff		RLH+SALC 95% cutoff	
n=	% Retained	Index	% Retained	Index
2000 Country A + 300 ATR	73%	1.28	70%	1.29
2000 Country A + 1800 ATR	68%	1.27	73%	1.21
2000 Country B + 300 ATR	47%	1.62	37%	1.44
4000 Country B + 600 ATR	46%	1.63	40%	1.51
4000 Country B + 1800 ATR	39%	1.69	35%	1.31
2000 Country B + 1800 ATR	33%	1.73	68%	0.88

In Country A, where we expect to have fewer “bad” respondents, RLH only and RLH+SALC remove about 30% of respondents. In Country B, where we expect more “bad” respondents, both methods remove about 50–60%. Both are able to detect this difference in data quality.

The one outlier is Country B, n=2000 + 1800 ATR. The RLH+SALC method removes two-thirds of respondents and actually makes differentiation worse (index score < 1). Our best explanation is that having an overwhelming amount of test data (90% in this case), as well as very poor quality real data (over half of the real respondents are also behaving badly), makes it difficult for the SALC model to identify truly “bad” respondents.

RLH Only versus RLH+LC

Comparing RLH only to the RLH+LC method, the LC approach retains more respondents, but achieves less differentiation. The result is more noise and less signal. This method also retains about the same number of respondents in both countries, failing to pick up the difference in data quality between the two.

	RLH only 95% cutoff		RLH+LC 95% cutoff	
n=	% Retained	Index	% Retained	Index
2000 Country A + 300 ATR	73%	1.28	88%	1.07
2000 Country A + 1800 ATR	68%	1.27	94%	1.03
2000 Country B + 300 ATR	47%	1.62	82%	1.25
4000 Country B + 600 ATR	46%	1.63	84%	1.19
4000 Country B + 1800 ATR	39%	1.69	89%	1.13
2000 Country B + 1800 ATR	33%	1.73	88%	1.08

Amount of ATR Data

We also notice that the amount of ATR data affects the proportion of respondents retained for the RLH only method in particular. This is most obvious in the Country B findings, where we see the amount of ATR data increase from 15% to 45% to 90% and the proportion of respondents retained decrease from about half of the Country B data to only a third of it. We suspect this is due to the fact the ATR and real data were run together for this model, effectively bringing down the RLH of the real data.

		RLH only 95% cutoff	RLH+SALC (95% cutoff)	RLH+LC (95% cutoff)
n=	ATR	% Retained	% Retained	% Retained
2000 Country A + 300 ATR	15%	73%	70%	88%
2000 Country A + 1800 ATR	90%	68%	73%	94%
2000 Country B + 300 ATR	15%	47%	37%	82%
4000 Country B + 600 ATR	15%	46%	40%	84%
4000 Country B + 1800 ATR	45%	39%	35%	89%
2000 Country B + 1800 ATR	90%	33%	68%	88%

When we separate the ATR data from the real data to determine the 95% RLH cutoff, the amount of ATR data no longer has a noticeable effect on the percentage of respondents retained. Additionally, separate runs are just as good as combined runs for differentiation, summarized by the index scores.

		RLH Only (combined runs) 95% cutoff		RLH Only (separate runs) 95% cutoff	
n=	ATR	% Retained	Index	% Retained	Index
2000 Country A + 300 ATR	15%	73%	1.28	75%	1.26
2000 Country A + 1800 ATR	90%	68%	1.27	75%	1.27
2000 Country B + 300 ATR	15%	47%	1.62	44%	1.69
4000 Country B + 600 ATR	15%	46%	1.63	45%	1.66
4000 Country B + 1800 ATR	45%	39%	1.69	43%	1.68
2000 Country B + 1800 ATR	90%	33%	1.73	42%	1.70

A more in-depth look at combined versus separate runs can be found in the appendix to this paper.

RLH+SALC Threshold

We also compared thresholds for the SALC models: the 95th percentile cutoff based on predicted probability from the regression; and the equal misidentification cutoff recommended by Hoogerbrugge and de Jong. Of the two approaches, we found that the 95% cutoff does a better job of removing "bad" respondents and improving differentiation.

	RLH+SALC 95% cutoff		RLH+SALC Equal misID cutoff	
n=	% Retained	Index	% Retained	Index
2000 Country A + 300 ATR	70%	1.29	82%	1.19
2000 Country A + 1800 ATR	73%	1.21	84%	1.14
2000 Country B + 300 ATR	37%	1.44	70%	1.21
4000 Country B + 600 ATR	40%	1.51	63%	1.27
4000 Country B + 1800 ATR	35%	1.31	67%	1.12
2000 Country B + 1800 ATR	68%	0.88	82%	0.97

The RLH+LC models were both inferior to the SALC models (data not shown).

Concept Test Data

In addition to the MaxDiff exercise, respondents also rated 3 of the 24 concepts on 5 KPI measures such as appeal, usage, distinctiveness, etc. We calculated the top 2 box percentages for each concept for each KPI and calculated standard deviations across the concepts. We indexed the standard deviation for the cleaned sample against the total sample (before cleaning), as we did for the MaxDiff scores. The values shown in the table are the average across the 5 KPIs. An index value greater than one shows improvement in differentiation.

	95% cutoff		95% cutoff	
n=	RLH Only (Combined)	RLH Only (Separate)	RLH+SALC	RLH+LC
2000 Country A + 300 ATR	1.30	1.26	1.33	1.02
2000 Country A + 1800 ATR	1.33	1.26	1.24	1.00
2000 Country B + 300 ATR	1.80	1.91	1.48	1.19
4000 Country B + 600 ATR	1.64	1.59	1.55	1.09
4000 Country B + 1800 ATR	1.70	1.64	1.51	1.07
2000 Country B + 1800 ATR	2.16	1.96	1.17	1.08

After removal of "bad" respondents using the RLH only and RLH+SALC methods, we find similar improvements in differentiatiation of concept test results to what we saw with MaxDiff scores. Both do a fairly good job, whereas the RLH+LC approach is, again, less useful. Real respondents who behave as random in the MaxDiff exercise also contribute to randomness in the concept testing data. Bad MaxDiff responders are also bad concept testers.

Final Consideration

Of the methods tested, RLH alone and RLH+SALC performed the best in terms of removing more respondents that behave as random and improving differentiation in MaxDiff and concept test scores. The rates of removal and improvements in differentiation achieved are similar between the two. The question that comes to mind is: Are the two methods identifying the *same* respondents as "good" and "bad"?

Thankfully, there is reasonably good agreement between RLH alone and RLH+SALC about which respondents are "good" and which should be flagged as random behaving.

n=	% Match
2000 Country A + 300 ATR	88%
2000 Country A + 1800 ATR	87%
2000 Country B + 300 ATR	71%
4000 Country B + 600 ATR	82%
4000 Country B + 1800 ATR	72%
2000 Country B + 1800 ATR	52%

Agreement is better in the Country A data than in Country B. Again, we posit that it is more difficult to identify "bad" respondents when the amount of poor quality data is greater. Our research indicates this may be more of an issue for the SALC method than RLH only. This is further confirmed by the Country B data with n=2000 and 1800 ATR cases, with only a 52% match between methods. Such a large amount of test data is also not helpful in this case.

Conclusions and Recommendations

Both RLH only and RLH+SALC perform similarly well in terms of removing random-behaving respondents and improving differentiation in MaxDiff and concept test results. To a large extent they flag the same people for removal. Given this similarity and the complexity of implementing the SALC approach, we are hard pressed to say that the SALC method is worth the effort. We also found that RLH+SALC may not work well with very poor quality data and a large amount of test data, while RLH only did not appear to be affected as much by these issues. However, if there is a desire to pursue the SALC approach, we recommend the 95% cutoff over the equal misidentification rule.

Running HB with real and ATR data combined and using the 95th percentile ATR cutoff, the number of ATR cases does have an impact on the number of real respondents retained after data cleaning—more ATR data leads to fewer respondents retained. Given this, it makes sense to run the ATR and real data as separate HB models.

While we looked at ATR sample sizes ranging between 15% and 90%, we found that 15% of the real data is adequate assuming you have a large sample. For studies with small sample sizes, consider n=300 or more for ATR data, so that the 95th percentile of

RLH can be reliably calculated. While arguably the more ATR data the better for a reliable RLH cutoff, if ATR data and real data are combined for the HB runs, or the SALC approach is to be used, there is some risk of the random data (noise) overwhelming the good (signal).

After much analysis, by multiple researchers, the recommendation by Orme is the simplest and most effective in terms of removing random-behaving respondents and improving differentiation in MaxDiff (and concept test) results.

Limitations and Considerations

There is something of a chicken and egg problem with this area of research—often the more stringent the cutoff, the better the model fit. But at what point are we removing people who just don't fit the model, rather than truly random behaving respondents or bots?

Although we don't know which method is "right," our research provides some indicators. Comparing data from two countries with known quality differences helped to identify some of the strengths and weaknesses of the different methods. The use of concept test scores also helped to validate the findings.

Orme and Chrzan both acknowledged that sparse MaxDiff/CBC data make it more difficult to distinguish random from conscientious responders. We have encountered this problem in another study where we used a sparse MaxDiff design. Using the 95% RLH cutoff, we flagged 43% of the real respondents as being "bad"; a 75% cutoff flagged 26%; and a 50% cutoff flagged 15%. Historically our sample from the same source and audience has been quite good, so in this case we elected to use the 75% cutoff.

RLH gives us one flag for "bad" behavior. As Orme noted, other possible flags include length of interview and straight-lining, as well as verbatim reviews. An approach could be used in which a respondent is removed only after failing on multiple fronts.

Practically, sample suppliers are open to providing extra data cases prior to data cleaning. If we send back the "bad" respondents, they can be removed from their data source. It is important for researchers to understand their sample source and audience, so that they know how much over-sample is needed and are able to assess whether their approach to identifying "bad" respondents is flagging a reasonable number of cases.

Jane Tang

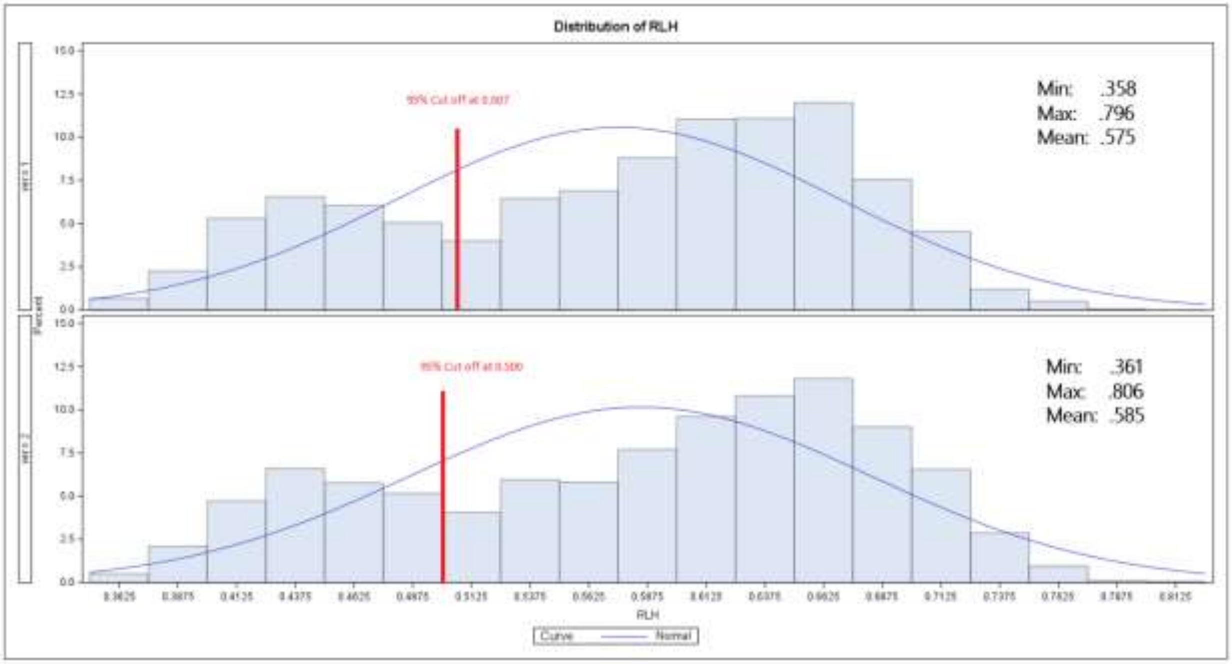
Mona Foss

Rosanna Mau

APPENDIX

A Further Look at Combined versus Separate Runs

We compared the RLH distributions for combined versus separate runs for Country A with 15% ATR data. In version 1, the 95% RLH cutoff was determined with an HB model in which the ATR data was run together with the real data and in version 2 it was run separate from the real data.



Country A, n=2000 (ATR n=300)

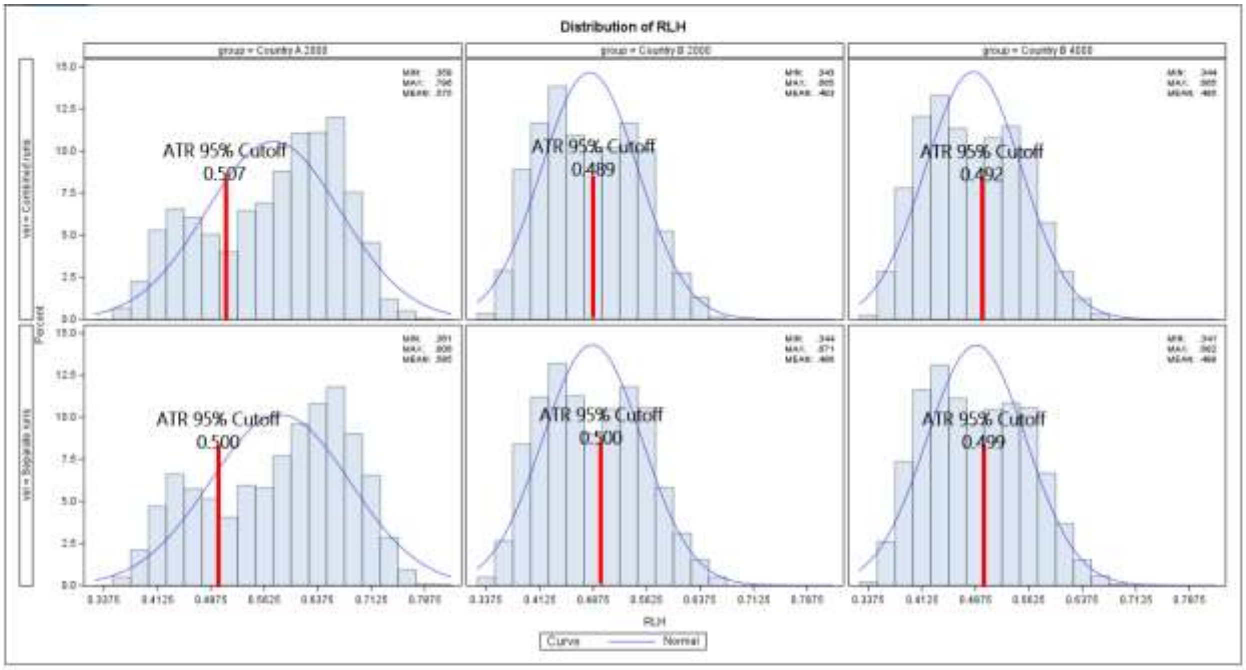
Note that in the combined run (ver 1), the full distribution of RLH for all real respondents shifts slightly to the left compared to the separate run (ver 2). In the combined run, the presence of the ATR data lowers the RLH of the real respondents.

However, the 95% cutoff is slightly higher (0.507 v. 0.500) in the combined run. HB borrowing in the combined run shifts the cutoff higher for the ATR respondents. This makes sense, since in Country A the majority of respondents are giving us thoughtful answers.

Regardless, there is still a 98% match between the two approaches in terms of respondents being flagged as “good” or “bad.” The combined run flags 35 additional respondents as “bad.”

In previous charts we saw that the greater the proportion of ATR data relative to real data, the greater the effect on this shift in cutoff when we use combined runs. Version 2 (separate runs) is stable regardless of how much ATR data we use.

When we compare these Country A findings to the Country B data sets also with 15% ATR, the story is somewhat different. Since the quality of the real data is quite poor, a much larger proportion of the real data are giving us random answers.



In the combined runs, the distributions of RLH among real respondents all shift to the left (regardless of country), but the cutoff point actually becomes lower in Country B, meaning that slightly more random-behaving real respondents are allowed to stay. With separate runs, the cutoff remains stable, as it should. Regardless, even with the differences in cutoff points, the two approaches flag the same people with a 97% to 98% match.

References

Chrzan, K and Halversen, C (2020) "Diagnostics for Random Respondents in Choice Experiments," Sawtooth Software Technical Paper. https://sawtoothsoftware.com/resources/technical-papers/diagnostics-for-random-respondents-in-choice-experiments

Hoogerbrugge, M. and de Jong, M. (2019), "Can we use RLH to assess respondent quality?" Sawtooth Software Conference Proceedings.

Orme, B.K. (2019) "Consistency Cutoffs to Identify 'Bad' Respondents in CBC, ACBC and MaxDiff" Sawtooth Software Technical Paper. https://sawtoothsoftware.com/resources/technical-papers/consistency-cutoffs-to-identify-bad-respondents-in-cbc-acbc-and-maxdiff

Mau, R., Tang, J., Helmrich, L., and Cournoyer, M. (2013) "Anchored Adaptive MaxDiff: Application in Continuous Concept Test," Sawtooth Software Conference Proceedings

UNCOMMON CHOICES: NOVEL APPLICATIONS OF CONJOINT ANALYSIS IN PRACTICE

CHRIS CHAPMAN
GOOGLE NBU (NEXT BILLION USERS)

ABSTRACT

This paper is a managerial review of four innovations in conjoint analysis practice. Although each case stands independently, the combination of methods demonstrates how experienced conjoint practitioners may extend both their breadth and depth to tackle larger and more strategic projects. The cases include improved market prediction, anticipation of competitive response, and the application of choice-based conjoint analysis for psychographic segmentation. The descriptions here are high level, while technical details and code are provided in associated whitepapers from previous Sawtooth Software Conferences.

INTRODUCTION

In this paper, I review and link a series of prior work to demonstrate how relatively incremental innovations in conjoint analysis contribute to a vast expansion of applications for choice modeling practitioners. These innovations were presented at previous years of the Sawtooth Software Conference, among other venues. I recap the methods here, discuss how they fit together, and add new market observations and practical reflections for practitioners.

There are four core methods discussed here:

1. *Adaptive choice-based conjoint*, and how it may obtain high quality estimates
2. Using *game theory* with conjoint analysis to predict competitive response
3. Finding optimal product portfolios in the presence of competition, with *genetic algorithms*
4. Using choice-based conjoint to find *psychographic consumer profiles* (segments)

My goal here is to give an approachable introduction to the methods, explaining why each may be of interest to practitioners. Technical details, including links to R code for some methods, are available separately in the *Proceedings* of prior Sawtooth Software Conferences (see references in each section). For a more general introduction to the typical applications and problems for conjoint analysis with technology firms and products, refer to Love & Chapman (2007) and Chapman, Love, & Alford (2008).

For purposes here, I assume general awareness of the core concepts of conjoint analysis (Orme, 2014), types of conjoint analysis surveys (CBC and ACBC), and the foundations of market simulations. My aim is not to explain every application in depth but to discuss how and why each is useful, and how the methods and cases build on one another.

FOUNDATION: GETTING THE BEST INDIVIDUAL ESTIMATES WITH ACBC (CASE 1)

The first premise in this paper is simple: if we obtain the best data at the individual level, we will have the most and best options for later analyses. Aggregate, group-level analyses are generally OK with relatively imprecise individual-level estimates. However, when we have better estimates, we can make better predictions and improve other approaches such as segmentation and—as I argue in the following two sections—competitive and portfolio modeling.

In many projects, I have found that Adaptive Choice-Based Conjoint (ACBC) performs well at finding individual-level estimates that are effective for market share prediction (Chapman et al., 2009; Johnson & Orme, 2007).

Case 1: Background. We had developed and planned to launch a consumer electronics product "P1," when a competitor announced a product "C1" that would compete very closely. Arguably, C1 offered a strictly superior set of features at the same price point as P1. Our business leaders wondered whether P1 had any realistic chance of success vs. C1 in the market.

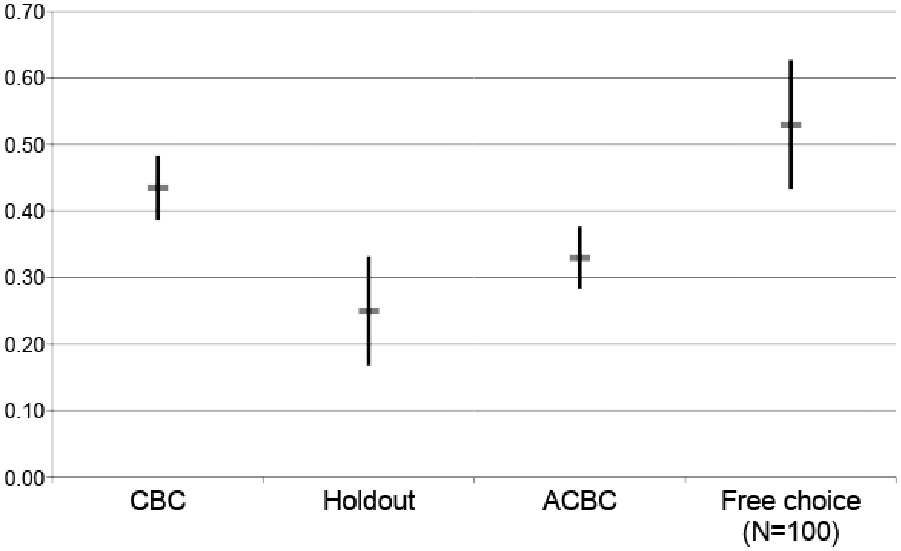
Case 1: Operational Research Question. The executive team decided that if our product achieved <25% estimated market share in a strict two-way head-to-head competition with competitor C1, then we would *cancel* the plan to launch P1. In my experience, it is rare for an executive team to give such a well-defined decision criterion. This reflects both the executive team's mature understanding of research, as well as previous experience and trust with this author's group and our research methods.

Case 1: Method and Results. We did not wish to make such a decision on the basis of a single method or estimate, and instead used 4 methods to estimate preference share for P1: (1) traditional choice-based conjoint (CBC); (2) a holdout task asked in CBC format, for the exact comparison of P1 and C1; (3) adaptive CBC (ACBC); and (4) a head-to-head offer for the respondent to choose one, either P1 or C1, presented in a richer, more descriptive style similar to retail marketing.

As shown in Figure 1, all 4 methods estimated P1 would attain preference share of at least 25%. We felt confident reporting that P1 would exceed 25% share, and thus there was no reason to cancel its launch. The firm subsequently released P1 as planned.

As detailed in the complete white paper (Chapman et al., 2009) we believed that ACBC provided the most credible estimate with a point estimate of 33% preference for P1 vs. C1. This was due to the greater amount of information collected in ACBC for each respondent, along with indicators of better internal consistency, such as the lack of attributes that showed level reversals. We reported the estimate of 33% to the executive team as our best prediction of consumer preference.

Figure 1: Estimated preference share for P1 vs. C1, using 4 methods.



How did it perform? Several months after the release of P1 and C1, we observed an actual market share of 34.6% for P1 vs. C1, compared to our estimate of 33%—a difference of less than 2% absolute, which was well within the confidence interval in ACBC market simulation.

Case 1: Notes for success. At the conference, we discussed whether this is an expected level of accuracy for ACBC. That is impossible to answer, but I would make a few points. First, this is not a unique or exceptional case in my experience. I have had the opportunity on only a few occasions to compare conjoint estimates to actual market data in any clear way. On each occasion, conjoint has performed well (see Chapman, Alford, & Love, 2009). However, I believe it is not the expected performance for the *first* occasion to conduct a conjoint analysis study for a product line. Such studies require iteration to learn about the attributes that matter, how to ask them, and how to tune aspects of the model such as brand effects. Notably, although we had conducted CBC on many occasions in this product space, this was our first application of ACBC.

Overall, my conclusion is that ACBC is capable of delivering the best individual-level estimates, and these may lead to the best-performing preference estimates. This comes at the cost of greater survey complexity and respondent time. Table 1 summarizes some of the considerations between CBC and ACBC.

Table 1: Brief comparison of considerations for CBC vs. ACBC.

	CBC	**ACBC**
Survey length	**Shorter**, 4–10 minutes	~2x as long, 8–15 minutes
Sample	Larger standard deviations ~2x sample needed for same precision	Better with **smaller samples**
Individual-level precision	Moderate, depending on number of tasks	High. Especially good when **segmentation** is desired
Experimental design	**Full control**; easy to control for information density, nested attributes, prohibitions, etc.	Less control; depends on the ACBC process
Non-compensatory features	Limited assessment; relies on design matrix	**Moderate to high** ability to assess (higher=longer)

In cases where precision is crucial—as in the present business case—I highly recommend considering ACBC, and to make this possible by shortening other aspects of a survey. In the next section, we will see an analysis that benefits from the highest-quality individual estimates.

Optimize: Respond to Competition Using Game Theory (Case 2)

A common question for any product manager considering an action is, “How will competition respond?” In this section, I consider a case where there was an opportunity to improve the feature set of a product, but at a higher cost of goods. Would it be worth it?

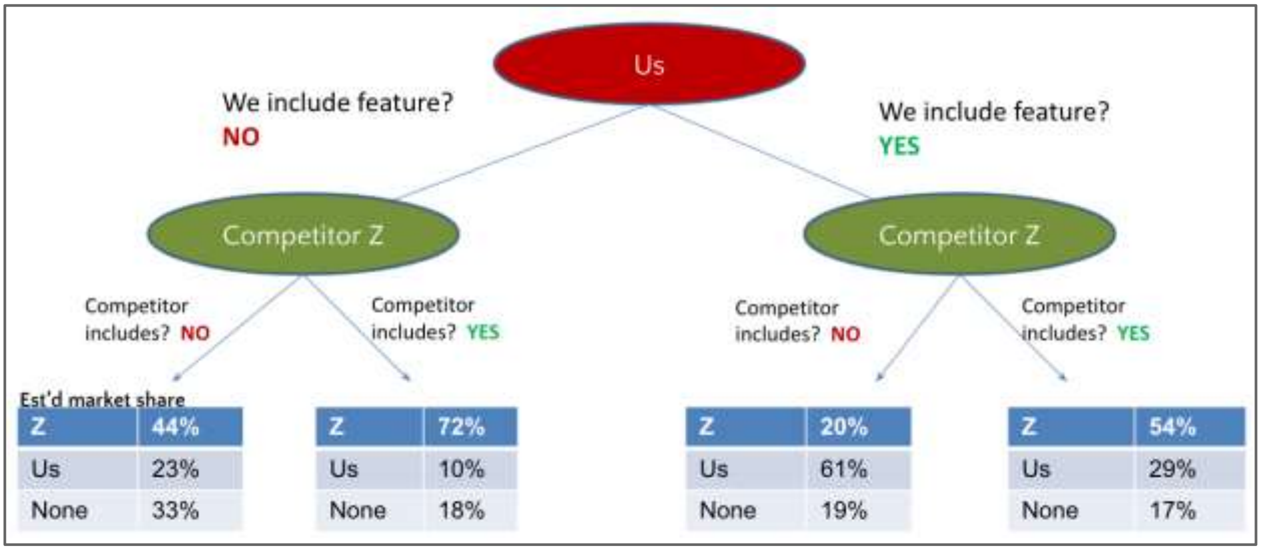
Case 2: Background. A consumer electronics product line “PL2” was presented with the opportunity to improve the nominal performance of a feature “F1” that was known to be highly salient for consumers. However, we knew that F1 would increase a product’s cost of goods and complexity of engineering, and believed that F1 might not yield any real improvement in the consumer experience. If our key competitor did not offer F1, and consumers ended up believing that F1 didn’t matter, then we would show lower profits for nothing. The executive presumption was that the competitor “Z” would *not* offer F1, and also that we should not; it was expected that our firm and Z would both benefit if F1 was in neither product line. This is structurally the same expectation as the so-called “prisoner’s dilemma,” where the overall best outcome is for both players to make the same choice of not “defecting”—in this case, adding F1—but only if both players make the same, independent decision (Myerson, 1991; Chapman & Love, 2012).

Case 2: Operational Research Question. Based on the success of prior analyses (such as Case 1 above), we expected that conjoint analysis was likely to yield a good answer as to the consumer value of F1 in the context of a competitive landscape. Unlike Case 1 above, in this case we were concerned with the effect on an entire product line, because F1 might be offered in multiple products and would be expected to shift preference within a product line.

We conducted a conjoint analysis survey and then estimated the net preference for preference of our line PL2 vs. the key competitor Z's product line "CL2" in four scenarios: that we offer F1 and they don't; that they offer F1 but we don't; that we both offer F1; and that neither of us offers F1 (for more on game theory and conjoint analysis, see Choi and Desarbo, 1993). In each case, we estimated the preference share for our line PL2 vs. their line CL2, in the presence of a "none" option estimate (see Karty, 2012, for discussion of the importance and difficulty of "none").

Case 2: Results. Figure 2 shows our model in the "extensive" format for a game theoretic analysis. First we consider what happens if we do *not* offer F1 in our line PL2, as shown in the two market simulation results on the left-hand side of Figure 2. The answer is that the competitor Z would see a large increase in preference share by offering F1—an increase from 44% preference to 72% preference vs. our line and the "none" option. Conclusion: if we don't offer F1, competitor Z will offer it.

Figure 2: Extensive form of the PL2 vs. CL2 competitive game for feature F1, with competition from brand Z.



Z	44%
Us	23%
None	33%

Z	72%
Us	10%
None	18%

Z	20%
Us	61%
None	19%

Z	54%
Us	29%
None	17%

Next, we consider what happens if we *do* offer F1, as shown on the right-hand side of Figure 2. We see that if we offer F1, then competitor Z should also offer F1—otherwise they obtain a share of 20% vs. a possible 54%. Conclusion: if we offer F1, then Z also will offer it.

Now, given that Z should offer F1, regardless of what we do, which choice would be better for us? Comparing the two "YES" paths for the competitor, we see that if they offer F1, then we should also offer it—that increases our preference share from 10% without F1 to 29% with F1. Conclusion: we should offer F1 if they do (and also if they do not).

Finally, what about the prisoner's dilemma? Comparing the left-most estimate in Figure 2 to the right-most estimate, we see this: if we both offer F1, we *both* expect to see market share increase. That is because the inclusion of F1 may pull in new purchasers vs. the "none" option.

The executive team was convinced by the data and analysis, and included F1 in our product line. *Were we right?* The product line share is difficult to estimate because the modeled portfolios for both us and the competitor rolled out over time. However, I can note the following: competitor Z initially did not offer F1 (in other words, the executive expectation about their likely action was correct), but our flagship product with F1 was highly successful, winning awards and achieving high sales. Z followed late and added F1 to their line many months later. In other words, the outcome closely matched the expectation from our game theoretic analysis, and our product line improved its competitive positioning. Finally, it turned out that feature F1 did in fact improve the user experience, which had been difficult to ascertain in advance.

Case 2: Keys for Success. As in Case 1, I cannot claim that such a successful outcome should be expected routinely, yet I also do not have examples of failure. Rather, my belief is that such success is attainable only if one has spent time to develop experience in a product area across multiple rounds of careful research and analysis. An important point in this case is that the "none" option turns out to be crucial: it is the part that breaks the prisoner's dilemma. That is a consideration that will need careful attention for such competitive modeling of a product line. Additionally, our experience in the space allowed us to add brand effects that produced better market share estimates, calibrated across multiple studies.

My recommendation from this case is simple: even basic game theory models may yield strong insight into likely competitive responses. In the complete paper, we consider a more complex model involving potential branding efforts (Chapman & Love, 2012). A corollary recommendation is this: don't bet against what customers are telling you. If customers want F1, then it's a good bet to give them F1, rather than attempting to outguess them.

This case considered a hand-crafted product portfolio vs. a competitor's expected portfolio, with regards to changing a single key feature. But what if we don't know what portfolio we should make, in light of many potential features? Can we get insight into an entire portfolio? The next section discusses how to optimize a portfolio using "genetic" search algorithms.

EXTEND: BUILD A PORTFOLIO USING GENETIC ALGORITHMS (CASE 3)

The previous sections discuss how to obtain improved estimates of product preference, and how to combine game theory with conjoint analysis to improve the definition and positioning for a product with regards to competition. In this section, I discuss how to generalize that analysis to an entire product line.

This extension answers three crucial questions. First is an obvious question: what is our optimal product line with respect to demand? Second, and closely related: how many products should we make? Finally, the third question may be more interesting: are there features that should be added to the product line?

Case 3: Background. For a consumer electronics product line (differing from Cases 1 and 2), the firm was making more than 20 SKUs, so many that engineering, marketing, and sales channels were excessively complex. The executive team wanted to know whether, and by how much, to reduce the product line size. One choice, of course, would be to cut the worst-performing products. However, consumer preference would shift around as the portfolio changes, and the popularity of products might not be related to their margin. For example, suppose a popular product has a lower margin. If we cut that product, would we lose sales to competition, or might we recapture them elsewhere in our portfolio, perhaps with currently less popular products that would have higher margin?

Case 3: Operational Research Questions. For purposes here, I will focus on two questions (see the complete writeup for others; Chapman & Alford, 2010). First, how many products should we make? We know that more products will always result in some additional share, but this should level off at some point where each additional product leads to a small incremental increase in share. In particular, if we find an optimal product portfolio for us (with the option to include a product portfolio for competitor Z), in the presence of the "none" option, how many products are needed to reach the point of strongly diminishing returns?

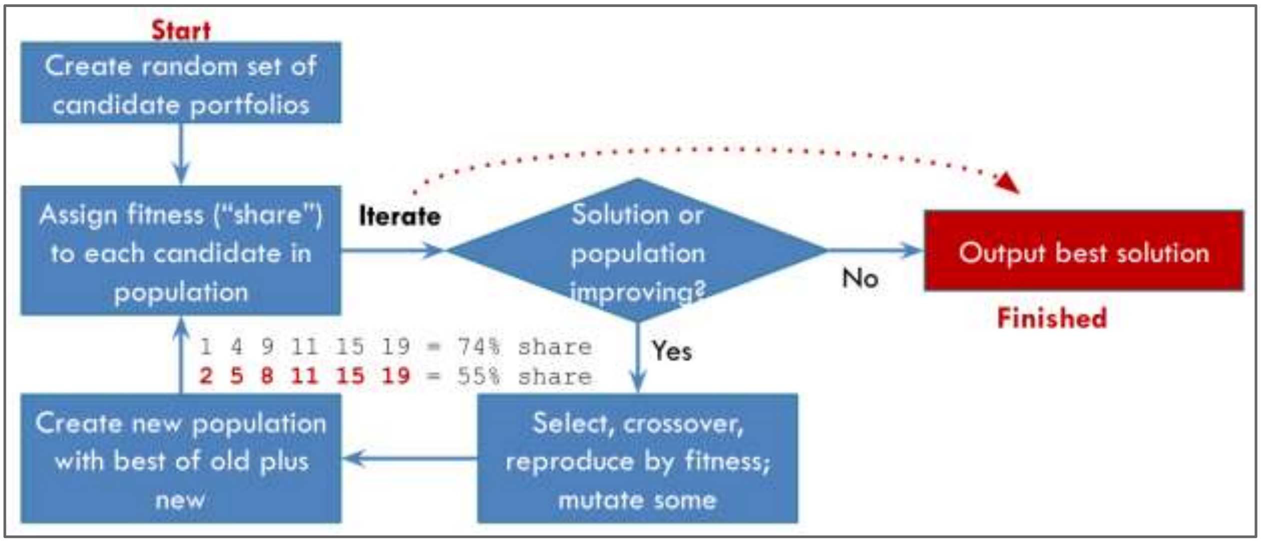
Second, is there a feature not currently offered in the product lines that often appears in optimal portfolios, and should be considered for addition to one or more products?

Case 3: Method. Our method to answer this question included four key elements. First, we wanted to obtain the highest-quality *data* (see Case 1 above). Not trusting a single method or survey, we opted to perform the analyses using two data sets from consumers that tested identical sets of attributes and levels; one using CBC and one using ACBC.

Second, we had to define an *outcome metric* to optimize. In this case, we optimized for the total preference *share* for any of our products—simply summed together, although one might instead optimize for revenue or profit, or a combination of metrics (Ferguson & Foster, 2013)—compared to the "none" option using randomized first choice market simulation. (Note that it is also easy to include competitors' products in the market simulation set, which then optimizes for share vs. competition, as in Case 2 above. In subsequent projects with this GA method, we have often done that.)

Third, there must be a *method to search* the product space. A genetic algorithm (GA) is an optimization method inspired by an analogy to evolutionary biology, in which best solutions are found by recombining parts ("genes") from prior, less optimal solutions. Belloni et al. (2008) demonstrate that GAs may achieve near optimal search of complex product spaces. Goldberg (1989) is an excellent technical guide to GAs. For conjoint analysis data, the genes represent product attributes and levels (i.e., product features, brands, and prices). Figure 3 presents a schematic representation of a GA approach to finding an optimal portfolio with conjoint analysis data. We implemented this in R using a standard GA library (Mebane & Sekhon, 2011; also R Core Team, 2021).

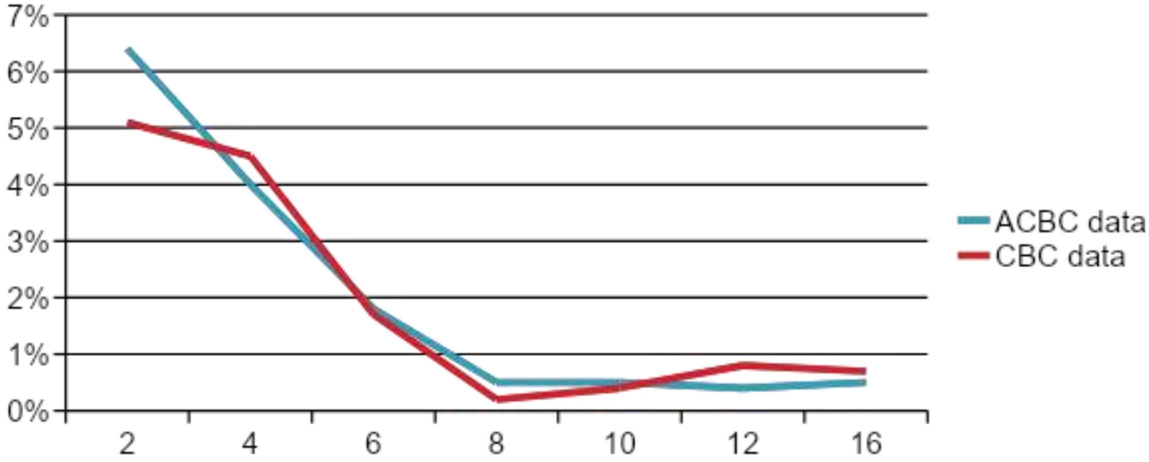
Figure 3: Outline of a genetic algorithm approach to find an optimal portfolio.


Start
Create random set of candidate portfolios
Assign fitness ("share") to each candidate in population
Iterate
Solution or population improving?
No
Output best solution
Finished
Yes
1 4 9 11 15 19 = 74% share
2 5 8 11 15 19 = 55% share
Create new population with best of old plus new
Select, crossover, reproduce by fitness; mutate some

The fourth and final part of the method is *iteration* over the space of potential portfolio sizes. The GA method is stochastic; any single "best" solution is one view, but we need more comprehensive insight into the *distribution* of possible solutions. We accomplished this by repeatedly finding GA solutions for each portfolio size. Specifically, we examined portfolio sizes from 1 to 20 total products in the product line; and ran 50 iterations of the model to find 50 unique near-optimal portfolio solutions for each size of product line. This was repeated for the two sets of data, from CBC and ACBC surveys.

Case 3: Results. Our first question involved optimal portfolio (product line) size. Figure 4 shows the incremental change in total portfolio preference share (sum of all products vs. "none") as the portfolio size increases. In results derived from both the CBC and the ACBC data, there was very little expected incremental gain in total preference when lines exceeded 8 products. Each additional product above 8 gains share primarily at the expense of other products within the portfolio, and only about 1% additional preference vs. "none." This suggested that we were making far too many products.

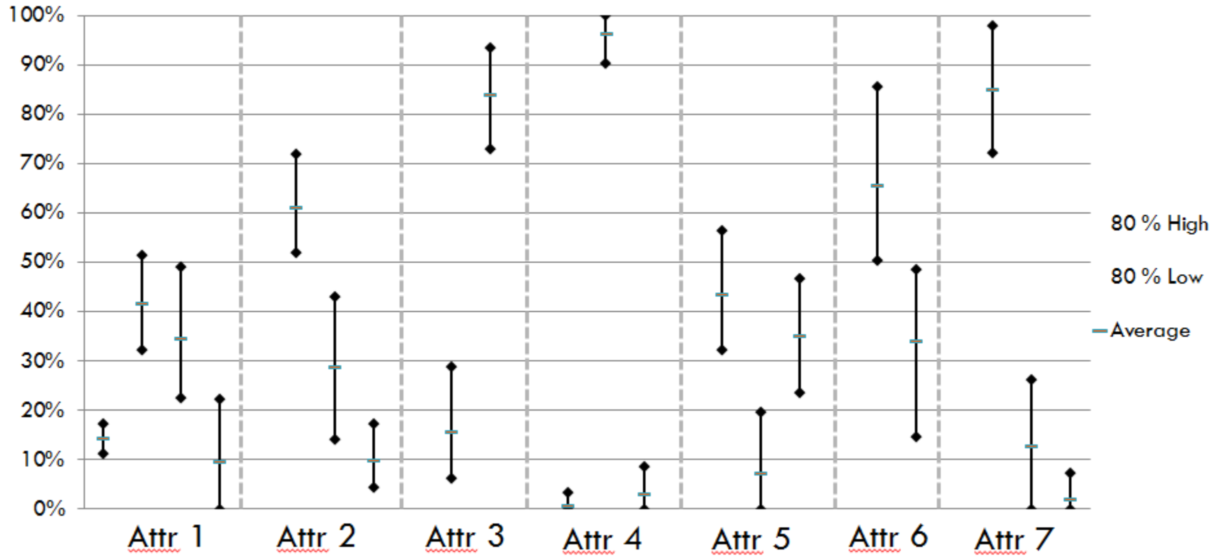
Figure 4: Change in total share of preference for the portfolio, for each incremental product, as the portfolio grows from 2 to 16 total products.



The second question was whether there was a key feature that should be in a product line but was not. Figure 5 shows the total preference share of the products in an optimal portfolio that include each CBC/ACBC attribute. For example, suppose there is an optimal line with 8 products, and 2 of those products include Attribute 2, Level (feature) 2, with shares of 10% and 5% each. We would say Attribute 2-2 thus has 15% total share.

One feature that was not currently present in any portfolio, for us or key competitors, was Attribute 2, Level 2. In Figure 5 we see that an optimal portfolio that includes Attribute 2-2 would have about 29% (range 14–43%) of preferred products with that feature. We concluded that this feature would be popular with consumers, as it should appear in roughly ⅓ of the products chosen in this product space, and we recommended consideration of including it in the product line.

Figure 5: Total preference share of products that include each feature (attribute level), in optimal portfolios.



What actually happened? First, the product line was shrunk. Similar to Case 2 above, it is difficult to determine the exact extent to which market results corresponded to our analysis, because the portfolio changes took significant time to achieve in the market, while other market variables continued to change. However, subsequent years did not see the firm return to a larger, expanded product line, so we can at least note that it was a stable strategy.

Second, because of its complexity, the product team decided not to include Attribute 2-2 in the product line, despite the expected consumer popularity. We view that as an analytic success because it was strongly considered, even if ultimately out of scope. We also note that 1.5 years later competitor Z introduced Attribute 2-2 in its product line. Several years later, it remains a core part of the product line for Z as well as other brands in the market. I would consider this also to be a success for our analytics team because we forecasted the desirability of that feature far in advance of its introduction, and thus anticipated a likely change in direction for consumers.

Case 3: Keys for Success. Similar to the preceding cases, the most essential factor in success for this case was high quality consumer data built on experience in this product space. We also conducted the analyses with data gathered through two methods, CBC and ACBC, and the agreement in analysis increased our confidence in the results. Another important aspect, similar to Case 2, is careful attention to the choice task design and the nature of the "none" attribute (Karty, 2012; Dotson et al., 2012; Huber, 2012; Chapman, 2013). Finally, it required innovation in method and a substantial degree of customized code (see the original whitepaper for more discussion, including availability of the R code; Chapman & Alford, 2010).

As I have noted already, I would caution an analyst not to expect such success in general; yet I also have no particular reason to expect less success. The case reflects our direct experience without "file drawer" selection issues. At the same time, our depth of experience in the product line, and our closeness to the engineering and executive teams were probably unusual, compared to a typical analyst's or external supplier's position. Such experience, communication, and trust no doubt contribute to the success of these kinds of strategic projects.

Understand: Develop Consumer Segments with Profile CBC (Case 4)

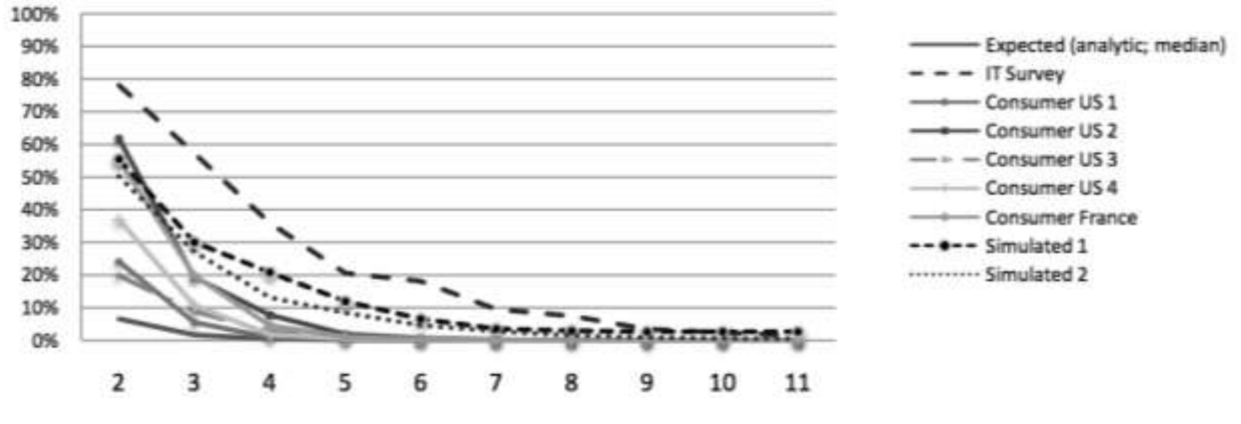
Cases 1, 2, and 3 demonstrated a logical progression in the depth of questions that one can answer with conjoint analysis data: good data leads to exceptional prediction of consumer preference; this may be used for competitive insight; and this can be broadened to answer highly strategic questions about an entire product line.

In the final Case 4, I turn to a different question and ask, if one becomes an expert at conjoint analysis (Orme & Chrzan, 2017), what other, highly novel applications might one tackle? In this case, we look at the problem of consumer segmentation, specifically the problem of building psychographic profiles that are sometimes known as "personas."

Case 4: Background. Marketers, designers, product managers, executives and other stakeholders often request descriptions of typical customers derived from segmentation or similar analyses. Often called personas or profiles, such descriptions are often produced by non-replicable, overfit, high-dimensional quantitative processes or qualitative methods (Chapman & Milham, 2006). Also, the descriptive dimensions are typically selected post hoc by the analyst and are not necessarily salient or particularly relevant to the user, and thus of little value for qualitative understanding.

It is typically impossible to say whether a persona is accurate, replicable, or descriptive of many—or any—customers (Chapman et al., 2008). Why? Because such a profile typically includes many descriptors; yet, as more descriptors are added, the proportion of users or customers who match the combined description will drop. This is one version of the "curse of dimensionality." Figure 6 shows this effect in several real and simulated data sets. In consumer data sets, a segment profile with 7 or more attributes is likely to be an exact match to *almost no one*.

Figure 6: Proportion of respondents who match the combination of values in a profile, as the number of variables in the description increases from 2–11 (Chapman et al., 2008).

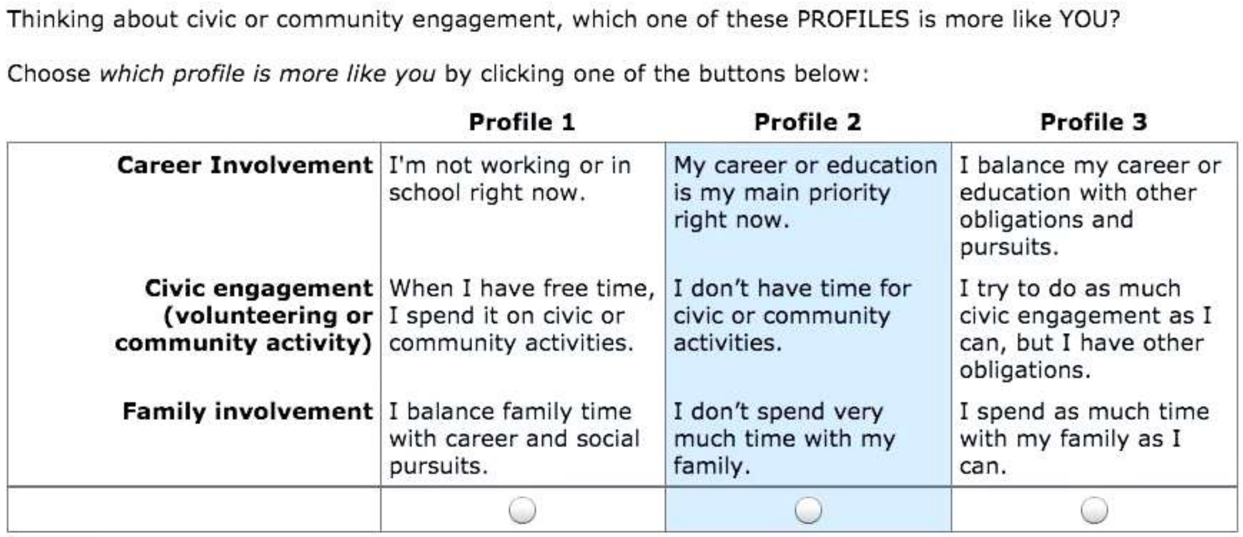


In Case 4, our research team was asked for user profiles, with regards to civic engagement, of adults in the US. We expected, on the basis of qualitative research, that there were many adults in the US who might be regarded as "interested bystanders," who are generally interested in civic information yet are not actively engaged in civic activities (Krontiris et al., 2015). The executive stakeholders wanted to understand these users better, and crucially, to know "How many interested bystanders are there?"

Case 4: Operational Research Question and Method. In light of the curse of dimensionality, our team sought a method that would do two things: (1) focus on identification from a *user's* point of view rather than our post hoc selection of variables, and (2) be less subject to the curse of dimensionality by assigning users probabilistically rather than categorically. We realized that choice-based conjoint analysis was exactly such a method (thanks to a suggestion from Greg Allenby, personal communication). We could find no prior example of psychographic segmentation using conjoint analysis, yet believed it was highly promising, and we differentiated it from general CBC by calling it "Profile CBC."

We selected attributes and levels to describe civic engagement, based on qualitative interviews conducted across the US. These were gathered into 8 attributes with typically 3 levels each, and used to form a CBC task asking users, "Which profile is more like you?" We found that the task was best constructed with 3 cards (3 profiles), without a "none" option, and with no more than 3 attributes shown in a partial profile format (Chrzan & Elrod, 1995; Patterson & Chrzan, 2003). Figure 7 shows an example task, as seen by a respondent (Chapman, Krontiris, & Webb, 2015).

Figure 7: Example task for Profile CBC, used for psychographic segmentation.



Thinking about civic or community engagement, which one of these PROFILES is more like YOU?

Choose *which profile is more like you* by clicking one of the buttons below:

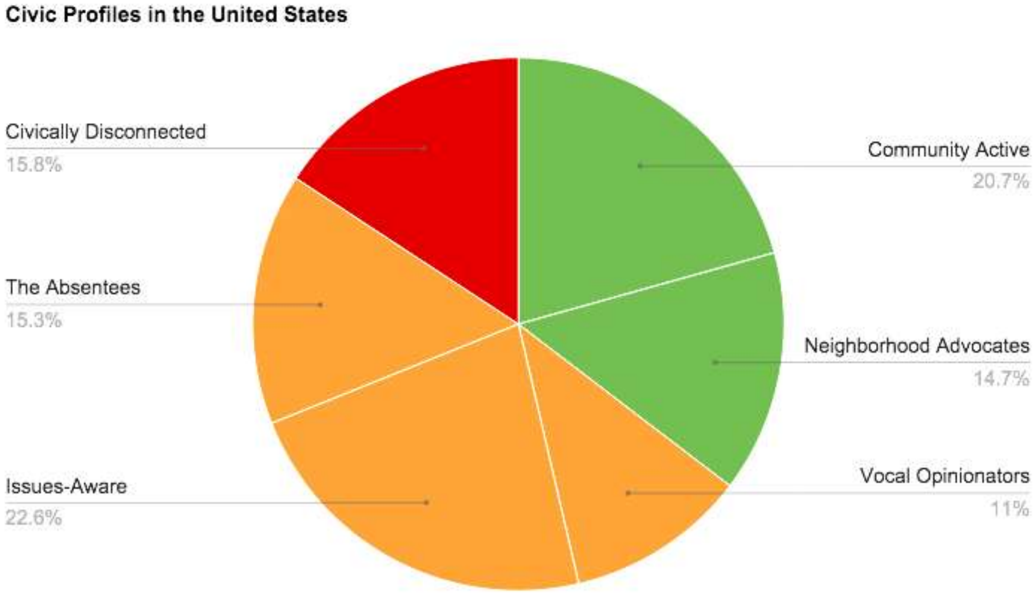
	Profile 1	**Profile 2**	**Profile 3**
Career Involvement	I'm not working or in school right now.	My career or education is my main priority right now.	I balance my career or education with other obligations and pursuits.
Civic engagement (volunteering or community activity)	When I have free time, I spend it on civic or community activities.	I don't have time for civic or community activities.	I try to do as much civic engagement as I can, but I have other obligations.
Family involvement	I balance family time with career and social pursuits.	I don't spend very much time with my family.	I spend as much time with my family as I can.
	○	○	○

Note that these design options are *unlike* the typical recommendation for CBC tasks that involve products. The choices of partial profile and omission of the "none" options would be unusual and not recommended for most product choice tasks, yet we found them to be warranted for a psychographic profile task. Otherwise, the task was too complex for respondents.

After collecting the CBC data on respondents' identification with the randomized profiles, we used latent class analysis (LCA; Sawtooth Software, 2021) to identify and size the potential segments for civic engagement.

Case 4: Results. Figure 8 shows the result from LCA analysis, in which an optimal solution identified 6 segments, with estimated sizes that ranged from 11–23% of the sample in each segment. In answer to the research question, we identified 3 segments—”absentees,” “issues-aware,” and “vocal opinionators”—who jointly comprised 49% of the sample and matched the concept of interested bystander. This answered the core executive question as to how many potential users we would target with our engineering and design efforts.

Figure 8: The psychographic segments and their sizing, as found by Profile CBC.



Additionally, we found that the segments were not only differentiated on the psychographic basis variables used for segmentation (i.e., the CBC attributes), but were also highly differentiated on other demographic and socioeconomic variables that were not used in the segmentation (for details, see Chapman, Krontiris, & Webb, 2015). In other words, the psychographic segmentation showed strong external validity with expected covariates and reported civic behaviors.

In short, the Profile CBC method achieved a useful result for psychographic segmentation. It answered the question of segment composition and size and did so on the basis of users' own reports about their identification rather than post hoc variable selection. Because the important identifiers were selected by respondents themselves, and applied through a probabilistic method (LCA, and conjoint analysis utilities in general), it was free from the most detrimental aspects of the curse of dimensionality. It also afforded the opportunity to explore options to recombine the segments on the basis of the underlying attitudes, i.e., to do “market simulation” in psychographic space.

Case 4: Keys for Success. There are two crucial aspects for Profile CBC: the attributes and levels must be appropriate, and respondents must be able to do the task. It is more difficult to specify appropriate attributes and levels than in a product-focused CBC, because the attributes are psychographic and attitudinal rather than a direct reflection of the product. The best approach would be a combination of qualitative and ethnographic research, plus any available baseline work on the prevalence and distribution of individual attitudes.

The other key concern is construction of the task such that it makes sense to respondents. In our trials, we found—again, *unlike* a product-focused CBC—that it was best to eliminate the "none" option and to force a choice among the cards; to limit it to no more than 3 cards at a time; and to use a partial profile approach with 3 attributes. Pre-testing with an in-person, think-aloud protocol is even more important than it is for product-focused CBC. For more design details and discussion of task format, see the full paper (Chapman, Krontiris, and Webb, 2015).

Conclusion

Taken together, the cases here demonstrate a "T-shaped" path for conjoint analysis practitioners, that extends both their breadth and depth of research offerings. With better respondent utilities (Case 1), consideration of competitive response (Case 2), and exploration beyond a product to an entire product line and brand (Case 3), skilled conjoint practitioners will be able to answer deeper, more important, and strategic questions accurately.

We have seen that conjoint analysis skills can also extend outside the product space, to conduct breadth research into consumers' attitudes, self-identification, and profiles (Case 4). I believe this "T" offers a much larger, more interesting, and more impactful space to inform decisions than simpler, single-point studies of product optimization (which remain highly important).

All four cases here demonstrated an arguably strong degree of external validity, ranging from successful market share prediction (Case 1), to anticipation of competitive responses (Cases 2 and 3), to alignment between psychographic and demographic variables (Case 4). I believe that such validation is possible only because of the attention to iteration within a product space, where the studies build on prior understanding of product features and attitudes. At the same time, the cases here—although novel—are not especially unique; I suspect that such success should be attainable in many product spaces, if one iterates and builds foundational knowledge.

In short, once one has developed expertise in product-focused CBC and the trust of executive sponsors, I highly encourage innovation! However, in such innovation, I strongly recommend building on existing best practices and methods (as with Case 1 and market simulation, or Case 2 and game theory) rather than creating novel statistical methods. In my experience, ad hoc tinkering with statistical methods is more likely to be a mistake than an advance.

If the reader is an R practitioner, I invite you to follow the development of our open source R package "choicetools" for conjoint analysis, MaxDiff, and related methods (Chapman & Bahna, 2019; Chapman, Alford, & Ellis, 2021). The package is in early development, and future releases will incorporate code for methods such as those in this paper (e.g., we will soon add code for the GA approach in Case 3, which is available separately today).

Finally, please consider sharing your cases, successes, and perhaps especially non-successes, as talks in future Sawtooth Software Conferences. None of the cases here would have been possible without the support, interchange, and inspiration my colleagues and I have received over the years from this community. It will make you a better practitioner!

Chris Chapman

REFERENCES

Belloni, A, Freund, AR, Selove, M, & Simester, D. (2008), Optimizing product line designs: Efficient methods and comparisons. *Management Science* 54:9.

Chapman, C. (2013). 9 things clients get wrong about conjoint analysis. In B. Orme (ed), *Proc. 2013 Sawtooth Software Conference*. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/sawtooth-conference-2013.pdf

Chapman, C., & Alford, J. L. (2010). Product portfolio evaluation using choice modeling and genetic algorithms. In B. Orme (ed), *Proc. 2010 Sawtooth Software Conference*. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/dc4cc660-880a-4e98-9049-32a3beb935d8.pdf

Chapman, C., & Alford, J. L. (2010, December). A genetic algorithm system for product line exploration and optimization. In *Proc. 2010 Second World Congress on Nature and Biologically Inspired Computing* (NaBIC) (pp. 268–273).

Chapman, C.N., Alford, J.A., and Ellis, S. (2021). choicetools. R package for conjoint analysis, maxdiff, and composite perceptual maps. Ver 0.9076. [R code]. Available at https://github.com/cnchapman/choicetools.

Chapman, C., Alford, J. L., Johnson, C., Weidemann, R., & Lahav, M. (2009). CBC vs. ACBC: comparing results with real product selection. In B. Orme (ed), *Proc. 2009 Sawtooth Software Conference*.

Chapman, C., Alford, J.L,. and Love, E (2009). Exploring the Reliability and Validity of Conjoint Analysis Studies. Poster presented at the 2009 Advanced Research Techniques Forum (ART Forum).

Chapman, C., and Bahna, E. (2019). choicetools: a package for conjoint analysis and best-worst surveys. Presented at UseR! 2019, Toulouse, France. Presentation; GitHub package.

Chapman, C., Krontiris, K., & Webb, J. (2015). Profile CBC: Using Conjoint Analysis for Consumer Profiles. In B. Orme (ed), *Proc. 2015 Sawtooth Software Conference*. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/40d2acfc-5474-48a6-a81a-aafbf4ea2686.pdf

Chapman, C. & Love, E. (2012, September). Game theory and conjoint analysis: using choice data for strategic decisions. In B. Orme (ed), *Proc. 2012 Sawtooth Software Conference*. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/1f6470c2-ddbb-4783-88f5-8c9b9b4d9c8f.pdf

Chapman, C., Love, E., Milham, R.P., ElRif, P., & Alford, J.L. (2008). Quantitative evaluation of personas as information. *Proc. Human Factors and Ergonomics Society (HFES) 52nd Annual Conference*, New York, NY, September 2008. http://goo.gl/4rLYEO

Chapman, C., Love, E., & Alford, J. L. (2008, January). Quantitative early-phase user research methods: hard data for initial product design. In *Proc. 41st Annual Hawaii International Conference on System Sciences* (HICSS 2008) (pp. 37–37). IEEE.

Chapman, C., Milham, R.P. (2006) The personas' new clothes: methodological and practical arguments against a popular method. *Proc. Human Factors and Ergonomics Society Annual Meeting*, 50:5, 634–636. SAGE Publications. https://goo.gl/szQ54E

Choi, S.C. and W.S. Desarbo (1993). Game theoretic derivations of competitive strategies in conjoint analysis. *Marketing Letters*, 4 (4), 337–48.

Chrzan, K., & Elrod, T. (1995) "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes." Presented at the 1995 Advanced Research Techniques Forum, Monterey, CA.

Dotson, J., Larson, J., Ratchford, M. (2012). Maximizing purchase conversion by minimizing choice deferral: examining the impact of choice set design on preference for the no-choice alternative. *Proc. 2012 Sawtooth Software Conference*, Orlando, FL. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/1f6470c2-ddbb-4783-88f5-8c9b9b4d9c8f.pdf

Ferguson, S., and Foster, G. (2013). Demonstrating the Need and Value for a Multi-Objective Product Search. *Proc. 2013 Sawtooth Software Conference*, pp. 275–304. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/sawtooth-conference-2013.pdf

Goldberg, D.E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

Huber, J. (2012) CBC Design for Practitioners: What Matters Most. *Proc. 16th Sawtooth Software Conference*, Orlando, FL, March 2012. http://goo.gl/ieqgMK.

Johnson, R.M., and Orme, B. K. (2007). A New Approach to Adaptive CBC (Whitepaper). Sawtooth Software, Provo, UT. https://sawtoothsoftware.com/resources/technical-papers/a-new-approach-to-adaptive-cbc

Karty, K. (2012). Taking nothing seriously: a review of approaches to modeling the "none" option. *Proc. 2012 Sawtooth Software Conference*, Orlando, FL.

https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/1f6470c2-ddbb-4783-88f5-8c9b9b4d9c8f.pdf

Krontiris, K., Webb, J., Krontiris, C., Chapman, C. (2015). Understanding America's "Interested Bystander:" A Complicated Relationship with Civic Duty. Technical report, Google Civics Research Workshop, New York, NY, January 2015. https://research.google/pubs/pub44180/

Love, E., and Chapman, C. (2007). Issues and Cases in User Research for Technology Firms. In B. Orme (ed), *Proc. 2007 Sawtooth Software Conference*. https://sawtoothsoftware.com/uploads/sawtoothsoftware/originals/dc4cc660-880a-4e98-9049-32a3beb935d8.pdf

Mebane, W.R., Jr., Sekhon, J.S. (2011). Genetic Optimization Using Derivatives: The rgenoud Package for R. *J. Statistical Software*, 42:11. http://sekhon. berkeley.edu/rgenoud.

Myerson, R. (1991), *Game Theory: Analysis of Conflict*. Harvard Univ. Press, Cambridge, MA.

Orme, B. (2014). *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research*, 3rd edition. Madison, WI: Research Publishers.

Orme, B, and Chrzan, K. (2017). *Becoming an Expert in Conjoint Analysis: Choice Modeling for Pros*. Sawtooth Software.

Patterson, M., & Chrzan, K. (2003). Partial Profile Discrete Choice: What's the Optimal Number of Attributes. *Proc. 2003 Sawtooth Software Conference*. https://goo.gl/uNqjTt.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sawtooth Software (2021). The Latent Class Technical Paper (version 4.8) (whitepaper). Sawtooth Software, Provo, UT. https://sawtoothsoftware.com/resources/technical-papers/latent-class-technical-paper

Comparing Predicted Automotive Purchase with Passive Geolocation Covariates to Actual Ownership Records

Edward Paul Johnson
Harris Poll
Marc Dotson
Brigham Young University

Introduction

With the rise of the mobile phone, researchers have now gained access to a deluge of data coming from the device. With appropriate consent, a panelist can allow researchers to study website visits (Zemla et al., 2015), apps used, social media connections (Kaur, 2016), calls made (Masso et al., 2019), purchases made (Montgomery et al., 2004), and locations visited (Hurwitz et al., 2010). Sometimes this data is anonymized and analyzed in aggregate (Sedayao et al., 2014); other times the behavioral data stream is the basis for triggering a survey (Poynter et al., 2014). In particular, location data has been analyzed extensively since the turn of the century (Asakura et al., 2014).

While the passive data stream coming from customers becomes increasingly important, it is unlikely to replace traditional survey research because each has its own unique advantages. While passive location tracking can inexpensively gather data on where people went with no recall error, it doesn't tell us much about why people went there. Survey data, while more expensive to collect per data point, is still needed to explore the motivations behind the actions. Survey data is also typically better structured. Experiments can be done to manipulate product characteristics such as in conjoint analysis (Orme, 2020). Survey data is also usually purposefully collected to be representative of a population rather than a byproduct of assisting people in obtaining directions (Baker, 2017).

When integrating or comparing survey and passive data it is important to consider both sets of data from a total error framework (Amaya et al., 2020). There are times when the two data sources will not agree due to recall error in the survey (Revilla et al., 2017). However, the passive data stream can contain both bad data and/or skewed data. Some examples include: missing data when a panelists turns off the meter, opt-in error around who agrees to be metered, processing errors around the data ingestion or transformation, and measurement error around false positives in measuring passive behavior. Lastly, sometimes essential aspects of model created are missing in the passive data which can lead to data incompatibility.

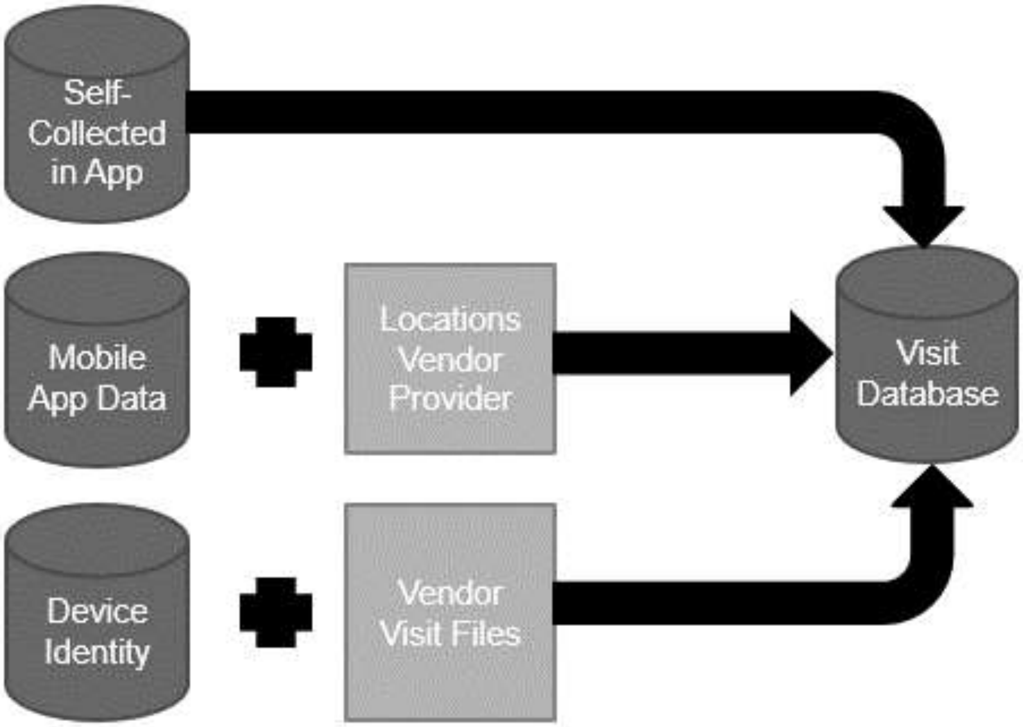
Data Sources

We examined the potential of combining passive location and survey data when predicting automotive purchases. To understand the differences between passive location data and survey data, we gathered both types of data among auto intenders in the United States. All sets of data came from Dynata, who gathered the appropriate permissions to integrate the data and supplied the data. We started with a sample frame of a subset from a non-probability panel who had given permission to have their location tracked; the results might not be generalizable to other populations. This purpose was not a research goal, so no attempt was made to weight the data by demographics to account for the coverage bias.

In May of 2018, we surveyed 784 panelists who intended to purchase an SUV in the next 6 months. Every panelist answered 12 standard CBC conjoint tasks on SUVs they would like to purchase with 8 attributes on each SUV. The survey also collected demographic data, price expectations and stated brand preference. Because we only invited those panelists who were tracked, we were able to append the number of visits to dealerships by brand for all respondents in the 6 months leading up to when they took the survey.

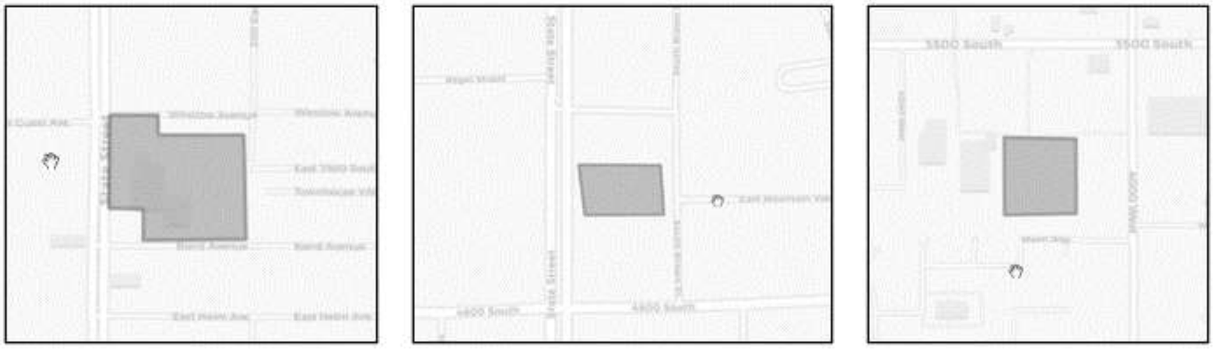
Dynata has multiple data streams for its passive geolocation visit data (Figure 1). The first is visits collected in their mobile QuickThoughts app directly from a point and radius for predetermined retail locations. We were not able to use this data stream because it is only available going forward after the locations are defined and we wanted 6 months of history in visiting dealerships. The second data stream is the historical latitude and longitude data that is collected passively through the app. These location "pings" can be intersected with polygons provided by a third party that define locations of interest. Lastly, Dynata will have mobile advertising ids (MAIDs) on those who have downloaded the QuickThoughts app and those ids could match to another vendor who specializes in passive visitation data. Almost all (95%) of the passive data in our study came from the second data stream while the remaining 5% came from the third data stream.

Figure 1



We also did some manual verification of the polygons used in the second data stream to confirm the accuracy of the data (Figure 2). The polygons and accuracy level of the geolocation data was reasonable, so we aggregated the data to provide a variable for every brand tested in the conjoint that described how many times the respondent visited a dealership affiliated with that brand in the past 6 months. If a dealership was affiliated with multiple brands the visit was counted towards each brand linked to that dealership.

Figure 2



Unfortunately, but not too surprisingly, the vast majority of the respondents did not visit a branded dealership in the past 6 months. Figure 3 shows the distribution of total visit data across all dealerships for the entire 6 months. The distribution of brands visited looked very reasonable for this audience (see Figure 4) with Ford, Honda and Chevy receiving the most dealer visits and luxury brands like Ferrari and BMW receiving the fewest.

Figure 3

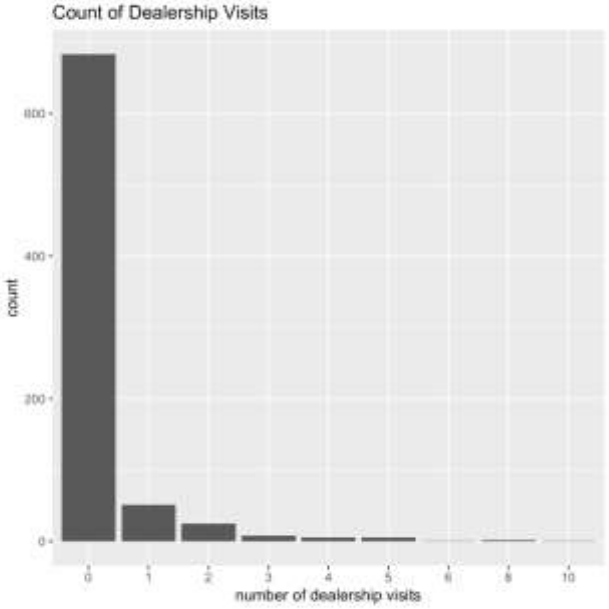
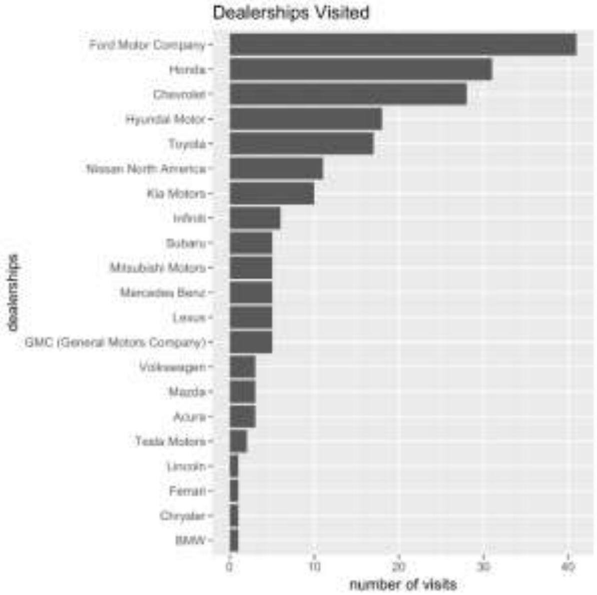


Figure 4



After building the models specified in the next section, we predicted vehicle choice for both in-sample and out-of-sample conjoint task holdouts. To validate the model we also wanted to compare to two more sources of data. In December of 2020 we recontacted those who completed the survey with a survey asking if they actually made a car purchase, and if they had, what type of car it was. We also reached out to a data provider to passively append automotive ownership records as a source of validation.

We were able to append vehicle ownership records for 187 of the respondents. These ownership records came from car service records linked to that person's name and address. Note that the passive data in this case was not directly what we were looking for (purchase records) and could have unknown bias in it based on the service shops that participated in sharing data with the vendor. We had 106 respondents complete the recontact survey (asking them if they made a purchase, and if they did, what was the make and model of the car they purchased). Both sets of data were sparse and, in the end, couldn't be used to validate the model because it was missing key characteristics of the car (such as the purchase price). It also included a lot of vehicles that were not included in the conjoint survey which focused on SUVs.

RESULTS

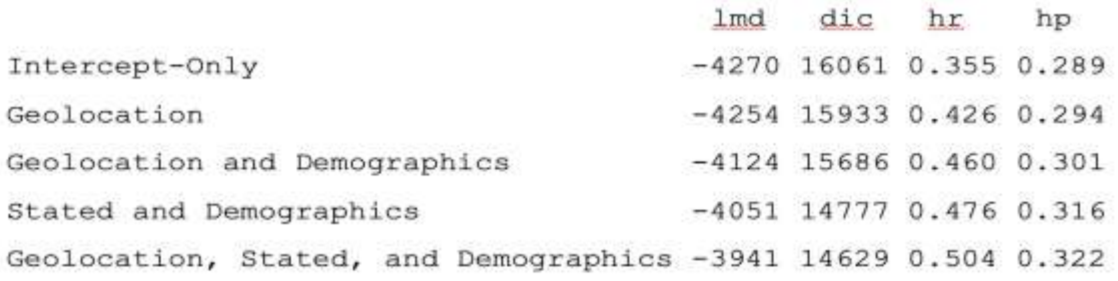
To determine the benefit of passive geolocation data compared to and in conjunction with stated preference data, we ran a series of hierarchical Bayesian multinomial logit choice models where the covariates included in the upper-level model differed. Using individual-level data as covariates in this way is natural since the upper-level in the hierarchy is a model of heterogeneity where the included covariates can help explain preference heterogeneity across respondents.

Additionally, when our interest is to use the model to make predictions for entirely new individuals (i.e., holdout respondents), the presence of covariates to inform drawing new sets of preference parameters (i.e., betas) for those new individuals can have a huge impact on predictive performance.

The results of the models are provided in Figure 5. The five models are named according to the covariates used in the upper-level model. The intercept-only model uses no covariates other than an intercept. The geolocation model uses the passive geolocation data, sparse as it is (see Figure 3), as covariates. The geolocation and demographics model uses both the geolocation data and the demographics data from the associated survey as covariates. The stated and demographics model uses a standard battery of stated preference variables along with the demographics data, both from the associated survey, as covariates. Finally, the geolocation, stated, and demographics model uses all the above as covariates in the model of heterogeneity.

Figure 5 also includes two sets of fit statistics: in-sample and out-of-sample (i.e., predicting to actual holdout respondents and thus reliant on the covariates being informative of preference heterogeneity as discussed previously). LMD is the log-marginal density and DIC is the deviance information criterion, both well-used measures of in-sample fit. HR is the hit rate while HP is the hit probability, both standard measures of out-of-sample fit, particularly in the presence of a classification model like a multinomial logit choice model.

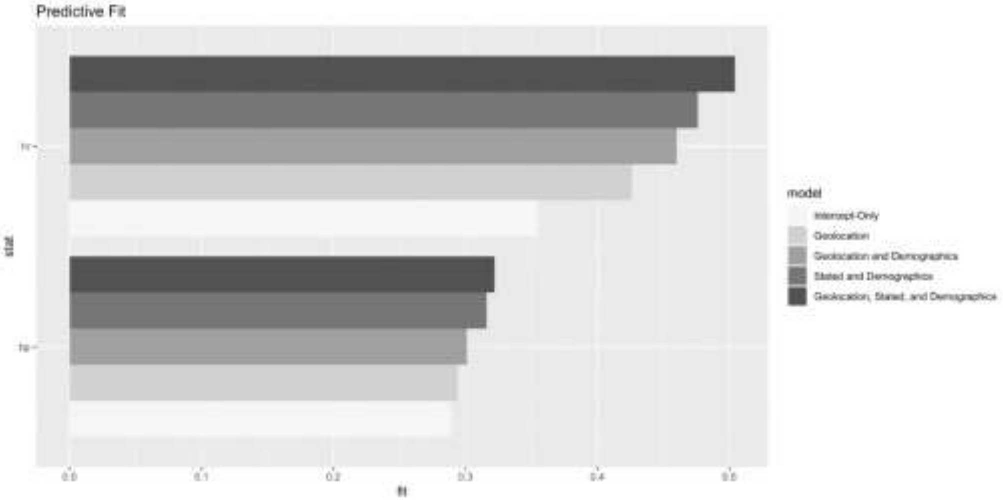
Figure 5



```
                                       lmd   dic    hr    hp
Intercept-Only                       -4270 16061 0.355 0.289
Geolocation                          -4254 15933 0.426 0.294
Geolocation and Demographics         -4124 15686 0.460 0.301
Stated and Demographics              -4051 14777 0.476 0.316
Geolocation, Stated, and Demographics -3941 14629 0.504 0.322
```

As we can see, all three sets of covariates are informative since including them all leads to the best model overall in terms of in-sample fit (closer to zero is better) and in terms of out-of-sample fit (larger is better). We especially care about predictive fit, and the improvement over sets of covariates becomes clearer as we chart the change in out-of-sample (i.e., predictive) fit in Figure 6.

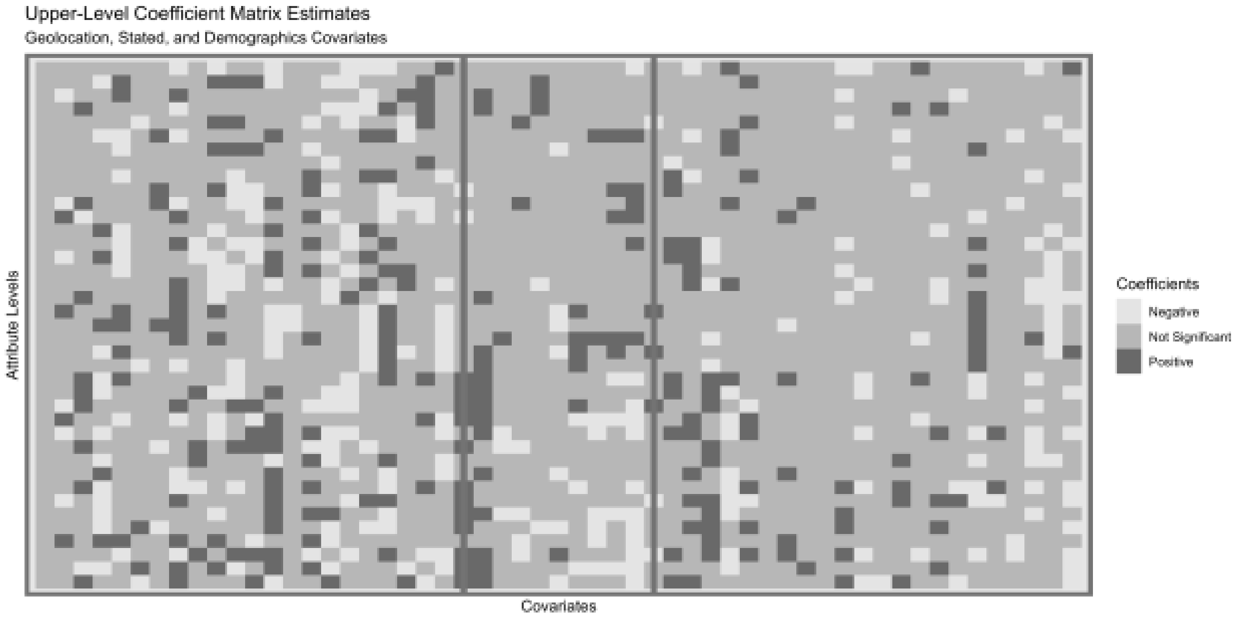
Figure 6



The geolocation covariates give us the biggest single boost in hit rate overall, while including all three of the sets of covariates provides the best prediction overall.

But what covariates matter? We can look at which marginal posterior distributions' 95% credible intervals don't include zero, which is a Bayesian approach to frequentist significance testing. Figure 7 is a heat map of the upper-level coefficient matrix estimates for the best-fitting model, colored according to "significance" and whether the impact is positive or negative on the resulting preference parameters.

Figure 7



Here we can confirm what we see in the predictive fit metrics: Across all three sets of covariates, we see covariates that have “significant” associations, both positive and negative, on the resulting preferences of individual consumers. In fact, the geolocation covariates appear to have the most “significant” associations with preference heterogeneity.

In particular, we can look at the marginal posteriors for the geolocation covariates to see the inferential impact of including these covariates in the model. Figures 8 and 9 plot the marginal posteriors of the attribute levels indicated on the y-axis for the brands in each facet. Note that in each figure we have filtered out the “insignificant” effects for clarity.

Figure 8

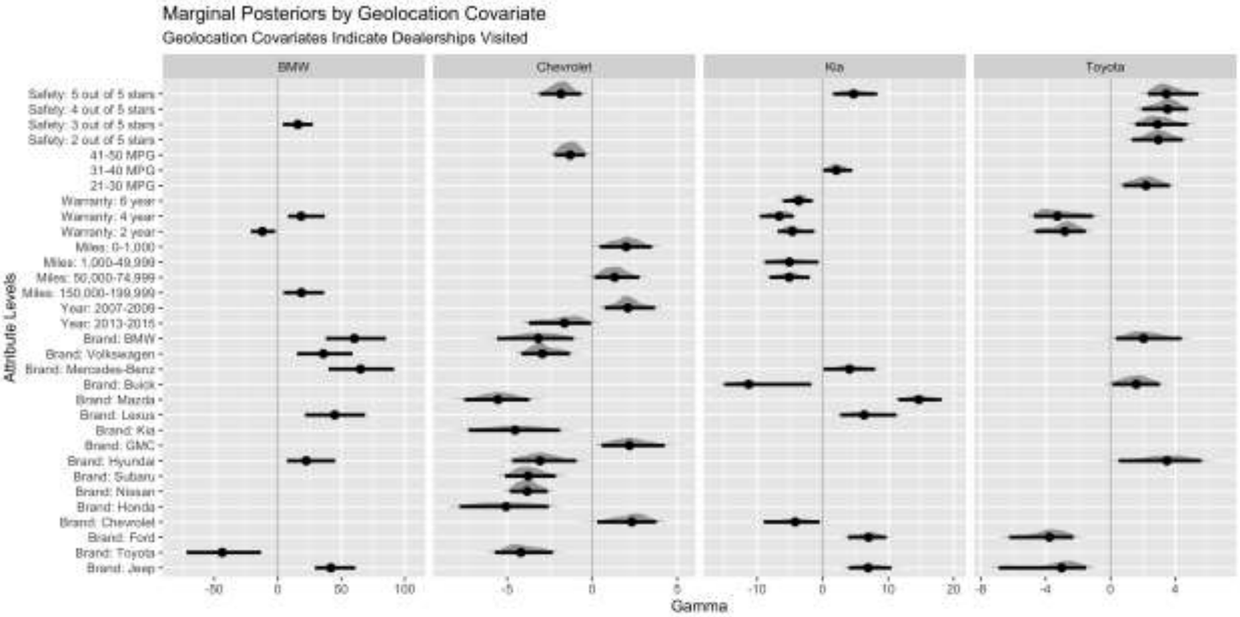
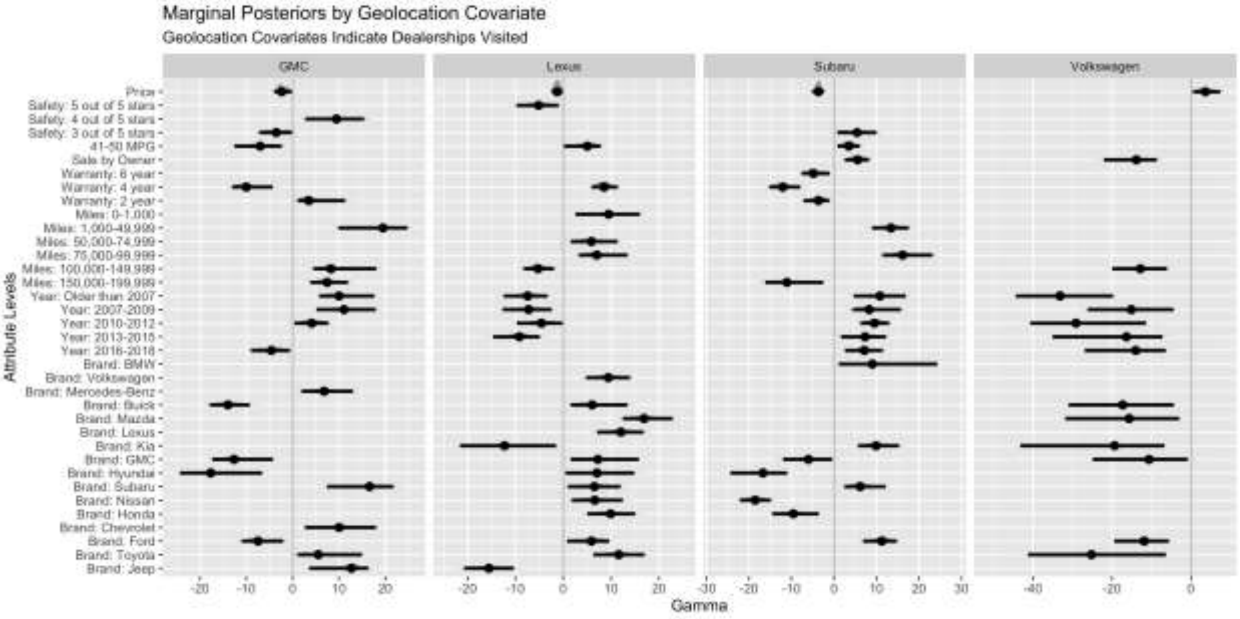


Figure 9



To illustrate how to read these plots, in Figure 8, we can see that those who have visited a Chevrolet dealership are less likely to pay attention to safety (i.e., the marginal posterior for Safety: 5 out of 5 stars is negative with reference to the holdout level of Safety: 1 out of 5 stars) and poor gas mileage (i.e., the marginal posterior for 41–50 MPG is negative with respect to the holdout level of 0–20 MPG) with a preference for older vehicles (i.e., the marginal posterior for Year: 2007–2009 is positive with reference to the holdout level of Year: 2016–Present). It appears that those who visit a Chevrolet dealer are most likely looking at the used car market and only for American-made vehicles (i.e., only Chevrolet and GMC has positive marginal posteriors). This illustration should help highlight the importance of having informative covariates in the model of heterogeneity when drawing betas for new respondents and thus performing well in terms of out-of-sample prediction.

Conclusion

Passive geolocation data can be effectively combined with survey data. When using the geolocation data, it is important to think critically about the data as it is not error free. Users should include data quality checks such as mapping the geolocation polygons on Google Maps, confirming visitation with survey data when possible, and aggregate brand-level checking. It is also important to confirm that the appropriate legal permissions have been given to use the geolocation data, especially if it is at the individual rather than aggregate level that is needed for conjoint covariates. Lastly, we recommend having a predefined use for the passive geolocation data as it is easier to digest the information that way rather than looking through all the possible locations that might correlate with variables of interest. Visits to retail locations are an obvious way to start, but we encourage researchers to explore other locations.

Passively collected geolocation data has the potential to substantially improve both inference and out-of-sample prediction, especially when used in conjunction with informative stated covariates. These covariates were able to substantially increase the predictive capability raising the out-of-sample hit rates from 35.5% to 50.4%. The significant covariates can also lead to insights about how the behaviors correlate with the car preferences. The model performance on holdout respondents supports the claim that geolocation data can be informative and a boost to inference and prediction.

We weren't able to effectively use repeated sampling and additional observational data to further validate the results. It is important when collecting either stated or passive validation data that the conjoint attributes are all provided on the validated data. When multiple sources of validation data are available, expect that they will not always agree and plan for what to do in the situations where they don't agree. Lastly, we recommend a robust initial sample size is also required as a good portion of survey respondents will not have validation data attached to it. Under the right circumstances and with the right format, we believe validation data could be useful in demonstrating the value of incorporating the passive data.

Edward Paul Johnson

Marc Dotson

REFERENCES

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total error in a big data world: Adapting the TSE framework to big data. Journal of Survey Statistics and Methodology, 8(1), 89–119.

Asakura, Y., Hato, E., & Maruyama, T. (2014). Behavioural data collection using mobile phones. Mobile technologies for activity-travel data collection and analysis, 17–35.

Baker R. (2017), “Big Data: A Survey Research Perspective,” in Total Survey Error in Practice, eds. Biemer P. P., de Leeuw E., Eckman S., Edwards B., Kreuter F., Lyberg L. E., Tucker N. C., West B. T., pp. 47–67, Hoboken: John Wiley & Sons, Inc.

Hurwitz, J. B., Wheatley, D. J., Zhang, K., & Lee, Y. S. (2010, September). Using Location History to Identify Patterns in Mobile Users’ Visits to Establishments. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 54, No. 5, pp. 507–511). Sage CA: Los Angeles, CA: SAGE Publications.

Kaur, S. (2016). Social media marketing. Asian Journal of Multidimensional Research (AJMR), 5(4), 6–12.

Masso, A., Silm, S., & Ahas, R. (2019). Generational differences in spatial mobility: A study with mobile phone data. Population, Space and Place, 25(2), e2210.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. Marketing science, 23(4), 579–595.

Orme, B. K. (2020). Getting started with conjoint analysis: strategies for product design and pricing research. 4th ed. Madison, WI: Research Publishers.

Poynter, R., Williams, N., & York, S. (2014). The handbook of mobile market research: Tools and techniques for market researchers. John Wiley & Sons.

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using passive data from a meter to complement survey data in order to study online behavior. Social Science Computer Review, 35(4), 521–536.

Sedayao, J., Bhardwaj, R., & Gorade, N. (2014, June). Making big data, privacy, and anonymization work together in the enterprise: experiences and issues. In 2014 IEEE International Congress on Big Data (pp. 601–607). IEEE.

Zemla, J. C., Tossell, C. C., Kortum, P., & Byrne, M. D. (2015). A Bayesian approach to predicting website revisitation on mobile phones. International Journal of Human-Computer Studies, 83, 43–50.

ENHANCE CONJOINT WITH A BEHAVIORAL FRAMEWORK

PETER KURZ
STEFAN BINNER
BMS - MARKETING RESEARCH + STRATEGY

BEHAVIORAL FRAMEWORK

As shoppers process information and act on it, they are not simple stimulus-response robots. Creating a behavioral framework prior to answering choice tasks therefore helps respondents select from choice tasks as if they were in a real purchase situation. If price and assortment changes are the focus of the research, it is particularly important to understand shopper perceptions of prices and values. Again, 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 apply nine standardized, binary "Behavioral Calibration Questions" regarding each respondent's individual shopping behavior for the focal category. Based on principles from behavioral economics, these questions help consumers recall their usual buying habits. "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.

BEHAVIORAL CALIBRATION QUESTIONS

We use the derived contextual information about each respondent's individual disposition towards brand and price knowledge (or lack thereof), past behavior, and perceptions within the category in our analysis. Retrieving a prior shopping situation and their individual dispositions helps consumers to make decisions in the following choice experiment. Currently, the set contains nine "Behavioral Calibration Questions" (semantic differentials) with respect to buying habits along three dimensions: brand, price, and innovation.

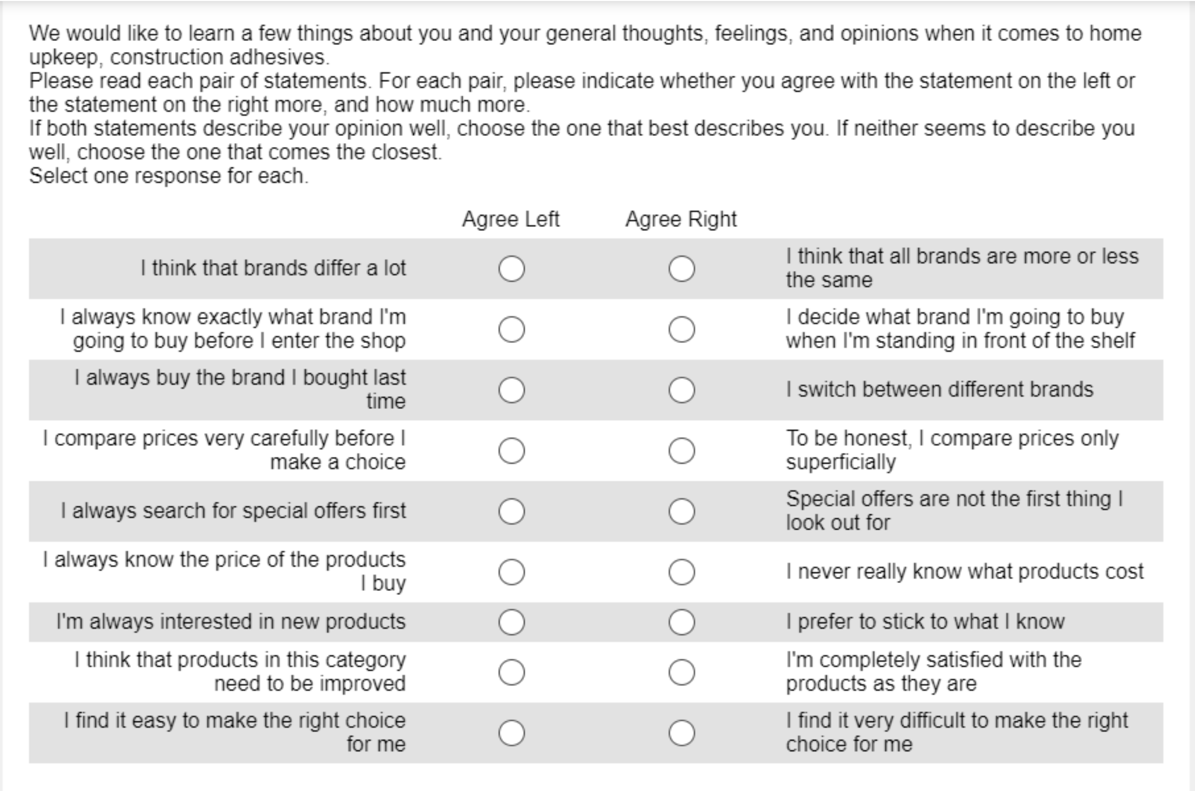
"Behavioral Calibration Questions" are used in our research context for several purposes:

- to establish a behavioral framework before respondents answer the choice tasks,
- as covariates in the hierarchical Bayes estimation process, and
- as segmentation/filter variables in the choice simulator.

Furthermore, we store the responses to the "Behavioral Calibration Questions" in a benchmark database to anchor further conjoint studies in the different product categories.

OUR STANDARD BEHAVIORAL CALIBRATION QUESTIONS

In each of our conjoint questionnaires, we combine the binary questions with our nine semantic differentials and ask respondents which of two statements (left or right) is more related to their last shopping trip.



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]

Of the nine semantic differentials, three are related to the “Role of Price,” three to the “Role of Brand,” and three to the “Role of Innovation.” This approach allows respondents to recall past behavior when buying a product in this category.

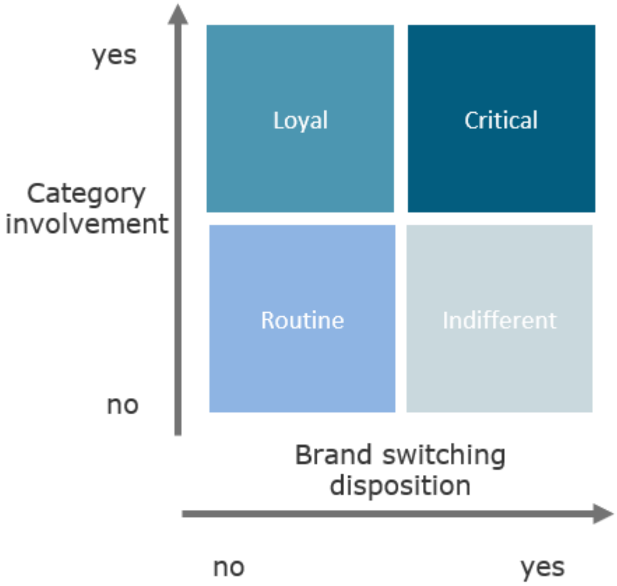
Over the years, we have adapted sets of nine semantic differentials to different product categories, as not all statements behave similarly in distinct shopping situations or categories. For example, when buying a new car, virtually no respondents would answer, “I never really know what the car I buy would cost.” In this situation, one needs a question such as “I never really know what the competitive brands cost; I more or less compare only within my preferred brand.” Such adaptations for each category are necessary to produce a valid framing of respondents from different target groups.

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

	FMCG	Media (Pay TV, VoD)	Mobiltelefone	PC_Notebook	Fernseher	Auto	Mobiltarife	Telekomtarife	Stromtarife	Girokonto	Spar-Anlagen	Krankenkasse	KFZ-Versicherung	Airlines	Pauschalreisen
1 - Markendifferenzierung	Die Marken unterscheiden sich stark voneinander	Die **Anbieter** unterscheiden sich stark voneinander	Die Marken unterscheiden sich stark voneinander	Die Marken unterscheiden sich stark voneinander	Die Marken unterscheiden sich stark voneinander	Die **Automarken** unterscheiden sich stark voneinander	Die **Anbieter** unterscheiden sich stark voneinander	Die **Anbieter** unterscheiden sich stark voneinander	Die **Anbieter** unterscheiden sich stark voneinander	Die **Banken** unterscheiden sich stark voneinander	Die **Anbieter** unterscheiden sich stark voneinander	Die **Krankenkassen** unterscheider sich stark voneinander	Die **Versicherungen** unterscheiden sich stark voneinander	Die **Airlines** unterscheiden sich stark voneinander	Die **Veranstalter** unterscheider sich stark voneinander
	Alle Marken sind mehr oder weniger gleich	Alle **Anbieter** sind mehr oder weniger gleich	Alle Marken sind mehr oder weniger gleich	Alle Marken sind mehr oder weniger gleich	Alle Marken sind mehr oder weniger gleich	Alle **Automarken** sind mehr oder weniger gleich	Alle **Anbieter** sind mehr oder weniger gleich	Alle **Anbieter** sind mehr oder weniger gleich	Alle **Anbieter** sind mehr oder weniger gleich	Alle **Banken** sind mehr oder weniger gleich	Alle **Anbieter** sind mehr oder weniger gleich	Alle **Krankenkassen** sind mehr oder weniger gleich	Alle **Versicherungen** sind mehr oder weniger gleich	Alle **Airlines** sind mehr oder weniger gleich	Alle **Reiseveranstalter** sind mehr oder weniger gleich
2 - Wichtigkeit der Marke bei der Kaufentscheidung	Ich weiß jedes Mal genau, welche Marke ich kaufen werde	Ich weiß **jetzt schon**, welchen **Anbieter** ich **wähle**	Ich weiß **jetzt schon**, welche Marke ich kaufen werde	Ich weiß **jetzt schon**, welche Marke ich kaufen werde	Ich weiß **jetzt schon**, welche Marke ich kaufen werde	Ich weiß **jetzt schon**, welche Marke ich kaufen werde	Ich weiß **jetzt schon**, welchen **Anbieter** ich **wähle**	Ich weiß **jetzt schon**, welchen **Anbieter** ich **wähle**	Ich weiß **jetzt schon**, welchen **Anbieter** ich **wähle**	Ich weiß **jetzt schon**, welche **Bank** ich **wähle**	Ich weiß **jetzt schon**, welchen **Anbieter** ich **wähle**	Ich weiß **jetzt schon**, welche **Krankenkasse** ich **wähle**	Ich weiß **jetzt schon**, welche **Versicherung** ich **wähle**	Ich weiß immer genau, welche **Airline** ich **buchen** werde	**Ich weiß genau, welchen Veranstalter ich wählen werde**
	Ich entscheide mich erst beim Kauf für eine Marke	Ich entscheide mich erst bei **Abschluss für den Anbieter**	Ich entscheide mich erst beim Kauf für eine Marke	Ich entscheide mich erst beim Kauf für eine Marke	Ich entscheide mich erst beim Kauf für eine Marke	Ich entscheide mich erst **unmittelbar vor dem Kauf**	Ich entscheide mich erst bei **Abschluss für den Anbieter**	Ich entscheide mich erst bei **Abschluss für den Anbieter**	Ich entscheide mich erst bei **Abschluss für den Anbieter**	Ich entscheide mich erst **kurz vor Abschluss für d. Bank**	Ich entscheide erst **kurz vor Abschluß über den Anbieter**	Ich entscheide mich erst **kurz vor Abschluss**	Ich entscheide mich erst **kurz vor Abschluss**	Ich entscheide mich erst **bei Buchung für eine Airline**	Ich entscheide mich erst bei **Buchung** für einen **Anbieter**
3 - Markenloyalität / Autopilot	Ich kaufe immer die gleiche Marke	Ich **wähle** immer den gleichen **Anbieter**	Ich kaufe immer die gleiche Marke	Ich kaufe immer die gleiche Marke	Ich kaufe immer die gleiche Marke	Ich wähle immer die gleiche **Automarke**	Ich **wähle** immer den gleichen **Anbieter**	Ich **wähle** immer den gleichen **Anbieter**	Ich **wähle** immer den gleichen **Anbieter**	Ich **wähle** immer die gleiche **Bank**	Ich **wähle** immer den gleichen **Anbieter**	Ich **wähle** immer die gleiche **Krankenkasse**	Ich **wähle** immer die gleiche **Versicherung**	Ich buche fast immer die gleiche **Airline**	Ich buche fast immer den **gleichen Reiseveranstalter**
	Ich wechsle zwischen verschiedenen Marken	Ich wechsle zwischen verschiedenen **Anbietern**	Ich wechsle zwischen verschiedenen Marken	Ich wechsle zwischen verschiedenen Marken	Ich wechsle zwischen verschiedenen Marken	Ich wechsle zwischen verschiedenen **Automarken**	Ich wechsle zwischen verschiedenen **Anbietern**	Ich wechsle zwischen verschiedenen **Anbietern**	Ich wechsle zwischen verschiedenen **Anbietern**	Ich wechsle zwischen verschiedenen **Banken**	Ich wechsle zwischen verschiedenen **Anbietern**	Ich wechsle zwischen verschiedenen **Krankenkassen**	Ich wechsle zwischen verschiedenen **Versicherungen**	Ich wechsle zwischen verschiedenen **Airlines**	Ich wechsle zwischen verschiedenen **Reiseveranstaltern**
4 - Preisvergleiche	Ich vergleiche Preise sehr genau, bevor ich kaufe	Ich vergleiche Preise sehr genau, bevor ich **abschließe**	Ich vergleiche Preise sehr genau, bevor ich kaufe	Ich vergleiche Preise sehr genau, bevor ich kaufe	Ich vergleiche Preise sehr genau, bevor ich kaufe	Ich vergleiche Preise sehr genau, bevor ich kaufe	Ich vergleiche Preise sehr genau, bevor ich **abschließe**	Ich vergleiche Preise sehr genau, bevor ich **abschließe**	Ich vergleiche Preise sehr genau, bevor ich **abschließe**	Ich vergleiche **Konditionen** sehr genau **vor Abschluss**	Ich vergleiche **Konditionen** sehr genau **vor Abschluss**	Ich vergleiche **Kosten** sehr genau **vor Abschluss**	Ich vergleiche **Kosten** sehr genau **vor Abschluss**	Ich vergleiche Preise sehr genau, bevor ich **buche**	Ich vergleiche Preise sehr genau, bevor ich **buche**
	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ich vergleiche **Konditionen** eher oberflächlich	Ich vergleiche **Konditionen** eher oberflächlich	Ehrlich gesagt vergleiche ich **Kosten** eher oberflächlich	Ehrlich gesagt vergleiche ich **Kosten** eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich	Ehrlich gesagt vergleiche ich Preise eher oberflächlich
5 - Sonderangebote	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach **Sonderkonditionen**	Ich suche immer zuerst nach **Sonderkonditionen**	Ich suche immer zuerst nach **Sonderkonditionen**	Ich suche immer zuerst nach **Sonderkonditionen**	Ich suche immer zuerst nach Sonderangeboten	Ich suche immer zuerst nach Sonderangeboten
	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	**Sonderkonditionen** sind nicht das Erste wonach ich suche	**Sonderkonditionen** sind nicht das Erste wonach ich suche	**Sonderkonditionen** sind nicht das Erste wonach ich suche	**Sonderkonditionen** sind nicht das Erste wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche	Sonderangebote sind nicht das Erste, wonach ich suche
6 - Preiskenntnis	Ich kenne genau die Preise der Produkte, die ich kaufe	Ich kenne die Preise der für **mich interessanten Angebote**	Ich kenne die Preise der für mich interessanten **Modelle**	Ich kenne die Preise der für mich interessanten **Modelle**	Ich kenne die Preise der für mich interessanten **Modelle**	Ich kenne die Preise der für mich interessanten **Modelle**	Ich kenne die Preise der für mich interessanten Tarife	Ich kenne die Preise der für mich interessanten Tarife	Ich kenne die Preise der für mich interessanten Tarife	Ich kenne genau die **Konditionen meines Girokontos**	Ich kenne genau die **Konditionen meiner Anlagen**	Ich kenne genau die **Kosten meiner Krankenkasse**	Ich kenne genau die **Kosten meiner Versicherung**	Ich kenne mich bei **Flugpreisen** gut aus	Ich kenne mich bei **Reisepreisen** gut aus
	Ich weiß nie genau, was die Produkte kosten	Ich weiß nie genau, was die **einzelnen Angebote** kosten	Ich weiß nie genau, was die **einzelnen Modelle** kosten	Ich weiß nie genau, was die **einzelnen Modelle** kosten	Ich weiß nie genau, was die **einzelnen Modelle** kosten	Ich weiß nie genau, was die **einzelnen Modelle** kosten	Ich weiß nie genau, was die **einzelnen** Tarife kosten	Ich weiß nie genau, was die **einzelnen** Tarife kosten	Ich weiß nie genau, was die **einzelnen** Tarife kosten	Ich weiß nie genau **die Konditionen meines Girokontos**	Ich weiß nie genau **die Konditionen meiner Anlagen**	Ich weiß nie genau, was **meine Krankenkasse** kostet	Ich weiß nie genau, was **meine Versicherung** kostet	Ich weiß nie genau, was die **Flüge** kosten	Ich weiß nie genau, was die **Reisen** kosten
7 - Interesse an neuen Innovationen	Ich bin immer an neuen Produkten interessiert	Ich bin immer an neuen **Angeboten** interessiert	Ich bin immer an neuen **Mobiltelefonen** interessiert	Ich bin immer an neuen **Computern** interessiert	Ich bin immer an neuen **Fernsehern** interessiert	Ich bin immer an neuen **Autos** interessiert	Ich bin immer an neuen **Tarifen** interessiert	Ich bin immer an neuen **Tarifen** interessiert	Ich bin immer an neuen **Tarifen** interessiert	Ich bin immer an **neuen Girokonto-Angeboten** interessiert	Ich bin immer an **neuen Anlage-Angeboten** interessiert	Ich bin immer an neuen **Angeboten** interessiert	Ich bin immer an neuen **Angeboten** interessiert	Ich bin immer an neuen **Flug-Angeboten** interessiert	Ich bin immer an neuen **Pauschalangeboten** interessiert
	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne	Ich bleibe lieber bei dem, was ich schon kenne
8 - Zufriedenheit mit Status Quo	Ich glaube, die Produkte müssen verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden	Ich denke, **das Angebot** muss noch verbessert werden
	Ich bin voll zufrieden mit den Produkten, wie sie sind	Ich bin voll zufrieden mit den **Angeboten**, wie sie sind	Ich bin voll zufrieden mit **Mobiltelefonen**, wie sie sind	Ich bin voll zufrieden mit den **Computern**, wie sie sind	Ich bin voll zufrieden mit den **Fernsehern**, wie sie sind	Ich bin voll zufrieden mit den **Autos**, wie sie sind	Ich bin voll zufrieden mit den **Tarifen**, wie sie sind	Ich bin voll zufrieden mit den **Tarifen**, wie sie sind	Ich bin voll zufrieden mit den **Tarifen**, wie sie sind	Ich bin voll zufrieden mit den **Girokonten**, wie sie sind	Ich bin voll zufrieden mit den **Anlagen**, wie sie sind	Ich bin voll zufrieden mit **Krankenkassen**, wie sie sind	Ich bin voll zufrieden mit **KFZ-Tarifen** wie sie sind	Ich bin voll zufrieden mit dem **Flugangebot**, wie es ist	Ich bin voll zufrieden mit dem **Reiseangebot**, wie es ist
9 - Sicherheit der Wahl vs. Ambivalenz	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht	Die richtige Wahl für mich zu treffen, fällt mir leicht
	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer	Die richtige Wahl für mich zu treffen, fällt mir schwer

Behavioral Calibration Questions

The first insight we derive from these nine questions is the identification of four respondent segments. Two semantic differentials, “brands differ a lot” and “always buy the same brand,” can be used to classify consumers according to “Brand Loyalty” and “Category Involvement,” thereby providing useful insights about the product category in general. Quantifying these different buyer segments is useful for identifying the best-performing strategies for products under investigation in the choice model.



This classification mostly refers to consumers’ attitudes towards brands. A consumer classified as “Indifferent” is not necessarily indifferent to other attributes. Segment names should not be taken too literally, as classifications represent only a rough outline of consumers’ personalities. For instance, a “Loyal” consumer may actually have a relevant set of two or three brands. What makes her a “Loyal” consumer is her self-perception as someone who sticks to her brand(s) (as opposed to consumers who are indifferent to brand), and her belief that the difference between her brand(s) and others really matters.

Loyal

highly involved, and committed to one favourite brand

Critical

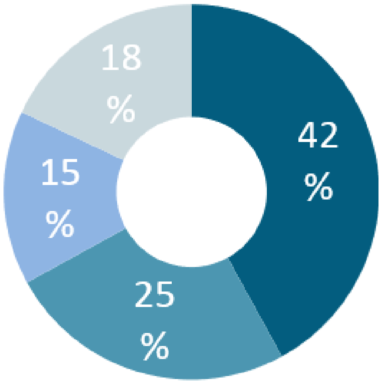
highly involved, but not committed to one favourite brand

Routine

uninvolved, habitually buying the same brand(s)

Indifferent

uninvolved and uncommitted



The figure above shows an example of the distribution of the four consumer types within the “laundry detergent” category. We see this as an initial blueprint for each category to begin interpreting the results of our choice models. Based on the benchmark from past studies, the client can easily determine how target consumers think about this category.

Experience with Behavioral Calibration Questions

Asking the nine “Behavioral Calibration Questions” before our choice exercise helps respondents to recall their behavior during their last shopping trip in a specific category. Therefore, we assume that the nine questions improve their decision-making process in the subsequent choice exercise, supporting a realistic answering behavior comparable to real shopping situations. Therefore, this approach helps generate more realistic data. Using the derived shopper classifications as segmentation variables in the choice simulator provides deeper insights into respondents’ preference structure. Based on our findings from numerous conjoint exercises, we learned that answering the nine questions results in better “Share of Choice” estimates as compared with conjoint exercises performed without the calibration questions. Furthermore, part-worth estimates, which include the “Behavioral Calibration Questions” as covariates, further improve share predictions against holdout samples (ensembles with the questions and other covariates offer marginal improvement in results).

Empirical Validation of Behavioral Calibration Questions

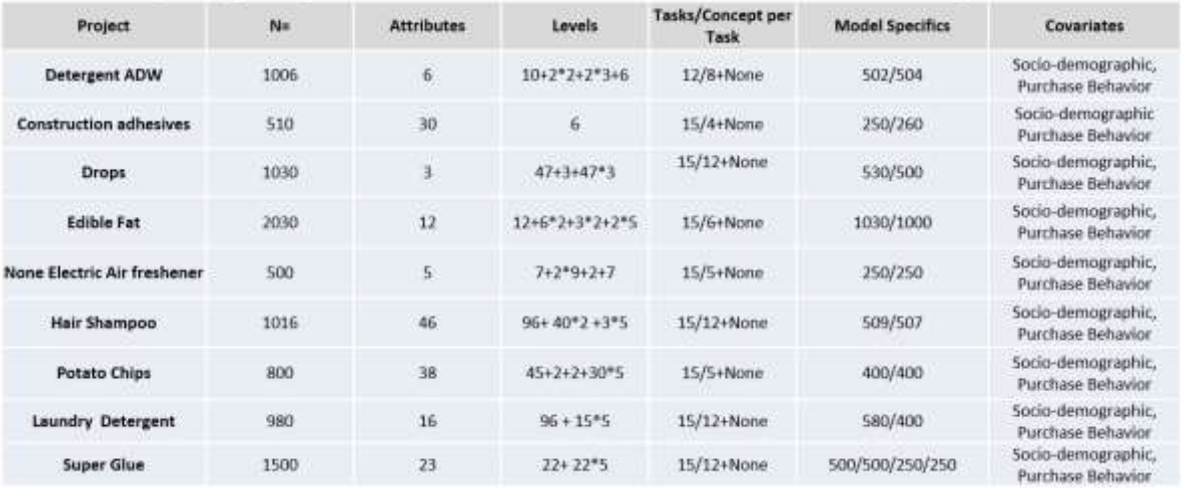
For validation purposes, we conducted nine empirical R&D studies over the last two years, in which we asked 50% of respondents the nine “Behavioral Calibration Questions” prior to answering the choice model, whereas the other 50% answered the choice model without being exposed to the semantic differentials prior to the choice tasks.

We addressed the following hypotheses in this paper:

- The framing offered by the "Behavioral Calibration Questions" results in improved answering behavior among our respondents, leading to part-worth estimates that are more stable and valid.
- Adding the answers to the "Behavioral Calibration Questions" as covariates in the HB estimation further improves the part-worth estimates.
- Using the questions as filter/segment variables in the choice simulator provides additional insights in the data as "Shares of Choice"; elasticities differ according to the derived segments based on roles of brand, price, and innovation.

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 varied 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.

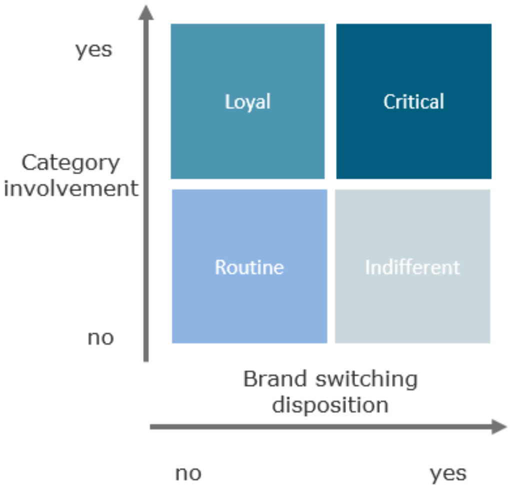
Only one study ("super glue") differed slightly from the others, as we conducted 4 sample splits to create an opportunity to validate the estimation samples with separate validation samples. (For the two estimation samples, n=500 interviews, and n=250 interviews for the two validation samples.) These 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.



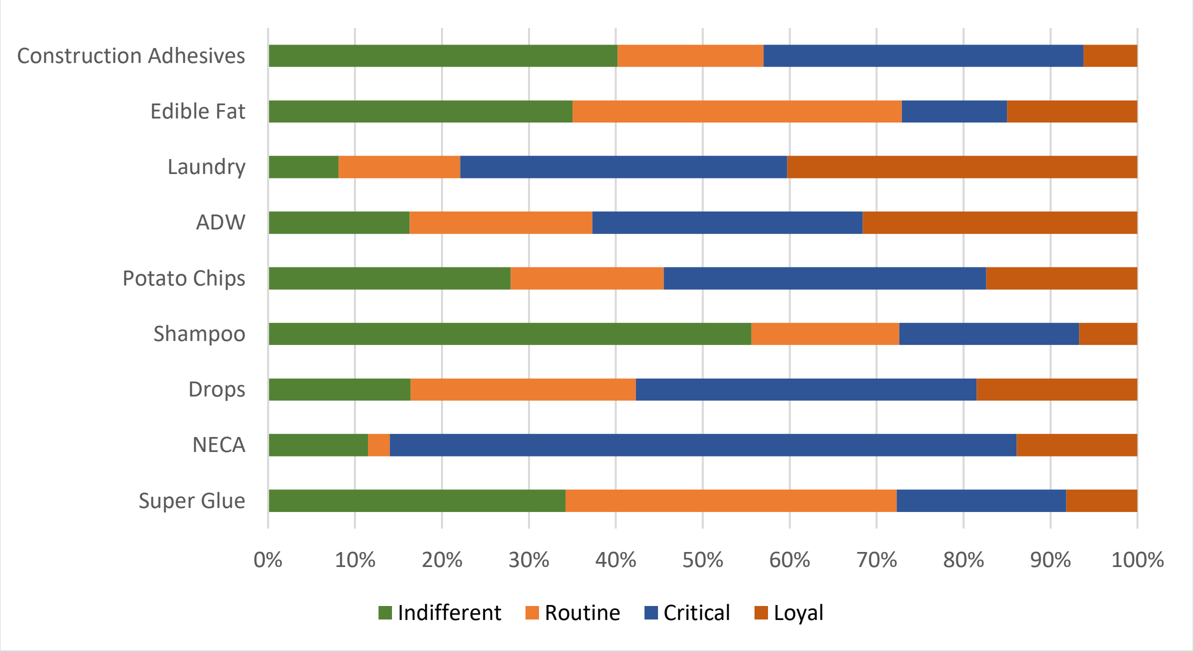
Project	N=	Attributes	Levels	Tasks/Concept per Task	Model Specifics	Covariates
Detergent ADW	1006	6	10+2*2+2*3+6	12/8+None	502/504	Socio-demographic, Purchase Behavior
Construction adhesives	510	30	6	15/4+None	250/260	Socio-demographic Purchase Behavior
Drops	1030	3	47+3+47*3	15/12+None	530/500	Socio-demographic, Purchase Behavior
Edible Fat	2030	12	12+6*2+3*2+2*5	15/6+None	1030/1000	Socio-demographic, Purchase Behavior
None Electric Air freshener	500	5	7+2*9+2+7	15/5+None	250/250	Socio-demographic, Purchase Behavior
Hair Shampoo	1016	46	96+ 40*2 +3*5	15/12+None	509/507	Socio-demographic, Purchase Behavior
Potato Chips	800	38	45+2+2+30*5	15/5+None	400/400	Socio-demographic, Purchase Behavior
Laundry Detergent	980	16	96 + 15*5	15/12+None	580/400	Socio-demographic, Purchase Behavior
Super Glue	1500	23	22+ 22*5	15/12+None	500/500/250/250	Socio-demographic, Purchase Behavior

Buying Habits and Involvement

The four segments derived from the "Behavioral Calibration Questions" have real potential to differentiate between categories and to identify promising strategies. For example, a significant proportion of "indifferent" consumers may have a larger effect on strategies for new product development, compared with a large share of "critical" or "loyal" consumers.



A comparison of the nine empirical studies shows that the different product categories have different compositions in terms of the four consumer segments. For example, in the "super glue" category, the study identifies an equal number of "Indifferent" and "Routine" consumers, whereas only a small proportion are "Loyal." In contrast, in the category "laundry," "Loyal" customers are by far the largest group, followed by the "Critical," "Indifferent," and "Routine" consumers. Furthermore, the "non-electric air freshener" (NECA) category has by far the highest share of "Critical" consumers. In the context of introducing new, innovative products, this category seems to offer many more opportunities as compared with the "super glue" category.



Behavioral Roles

The three behavioral roles represent a second possible usage of the nine calibration questions to understand the behavior of the respondents during shopping trips in a particular category. We found these roles helpful for interpreting the “Share of Choice” from simulations. They allow deeper insights into the reasons respondents behave differently in their choices.

The three roles derived from the “Behavioral Calibration Questions” are:

- Role of Price
- Role of Brand
- Role of Innovation

Each is represented by three semantic differentials that ask (in a binary manner) whether the left or the right statement better corresponds to the respondent’s last shopping trip in that category.

The following results, based on nine empirical studies, demonstrate the diversity of behavior within the different categories.

ROLE OF PRICE

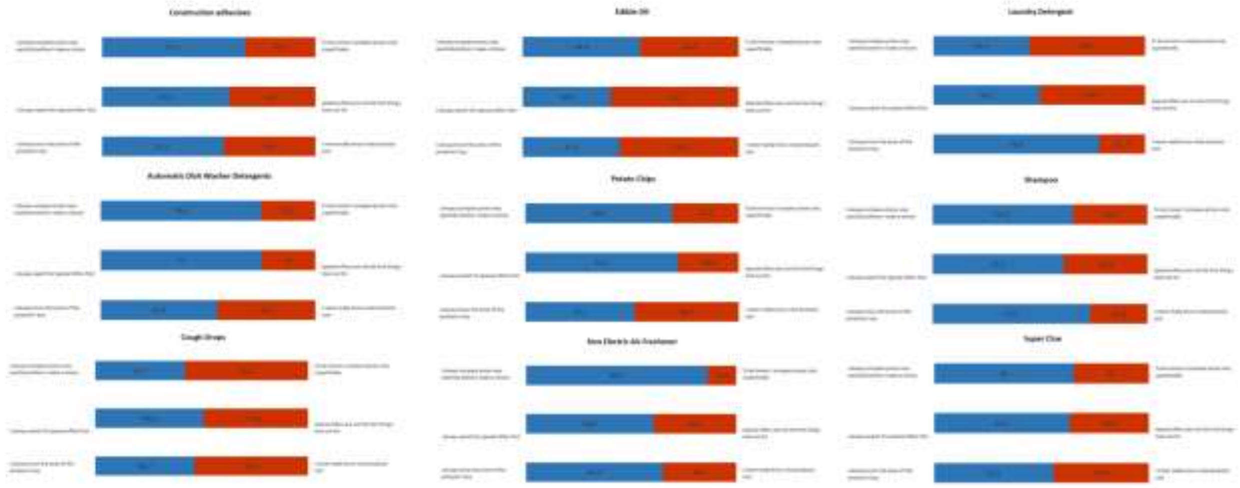
The “Role of Price” (RoP) is based on the following three semantic differentials:

“I compare prices very carefully before I make a choice”
vs
“To be honest, I compare prices only superficially”

“I always search for special offers first”
vs
“Special offers are not the first thing I look out for”

“I always know the price of the products I buy”
vs
“I never really know what products cost”

The following table shows how differently consumers behave when buying within these nine categories:



The number of consumers who always compare prices carefully varies between 41.9% (“cough drops”) and 86.4% (“NECA”). In the “edible oil” category, 40.8% are looking for special offers first, compared with 75% in the “automatic dish washer detergent” (“ADW”) segment. Price knowledge varies between 42% (“edible oil”) and 78% (“laundry detergent”).

Such differences in consumer behavior are useful for interpreting results from choice models. For example, in the “cough drops” category, a price increase is more likely to be accepted, given that 58% of the consumers do not compare prices. In contrast, only 15% do not compare prices in the “NECA” category, so price increases could have a much higher impact on preference shares.

Role of Brand

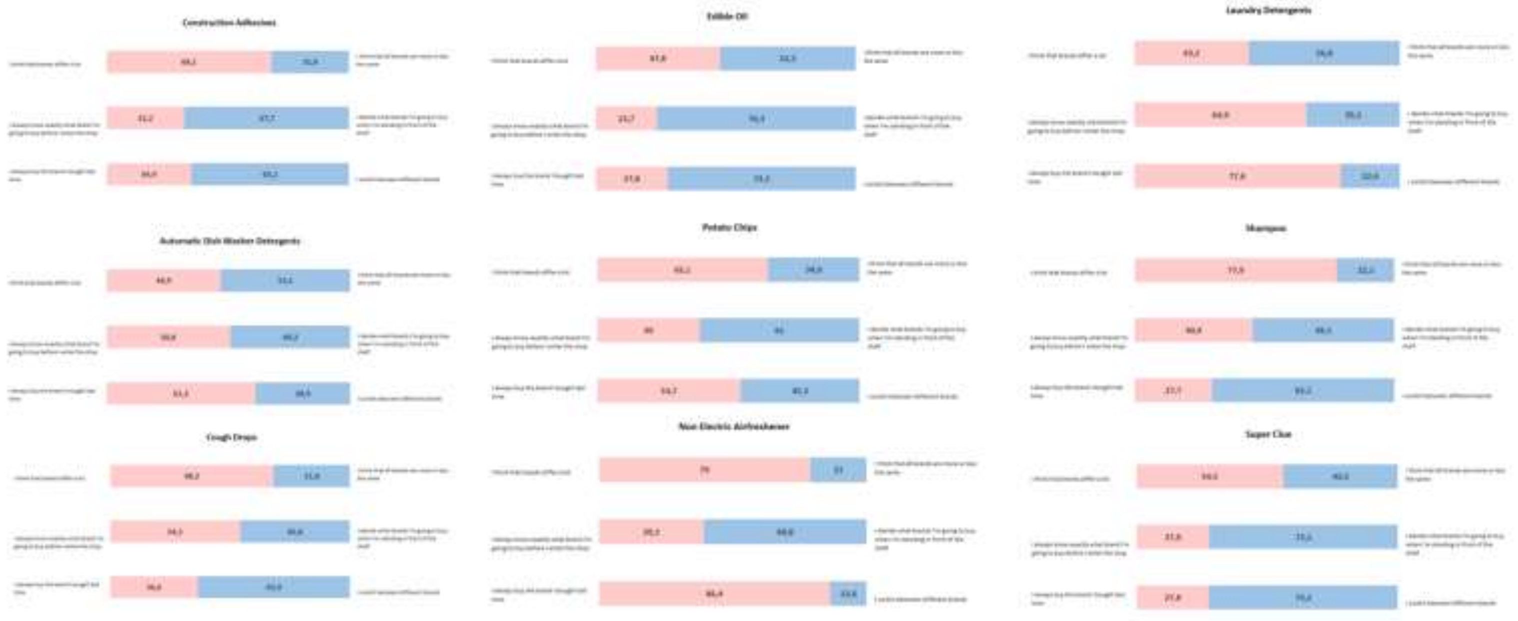
The “Role of Brand” (RoB) is represented by following differentials:

“I always buy the brand I bought last time”
vs
“I switch between different brands”

“I think brands differ a lot”
vs
“I think that brands are more or less the same”

“I always buy the brand I bought last time”
vs
“I switch between different brands”

These three differentials provide insights into the RoB, thus deepening understanding of consumers’ behavior in this regard. This approach provides further insight when interpreting simulations based on choice models.



Again, there are significant differences between the segments: The brand-switching attitude varies between 21% (“NECA”) and 56.8% (“laundry detergents”), representing a significant difference when a company aims to “introduce a new brand” into a category. Because 72% of “super glue” customers think that all brands are more or less the same, compared with only 14% of “NECA” customers, it seems that having a strong brand has more equity in the “NECA” category as compared with “super glue.”

Role of Innovation

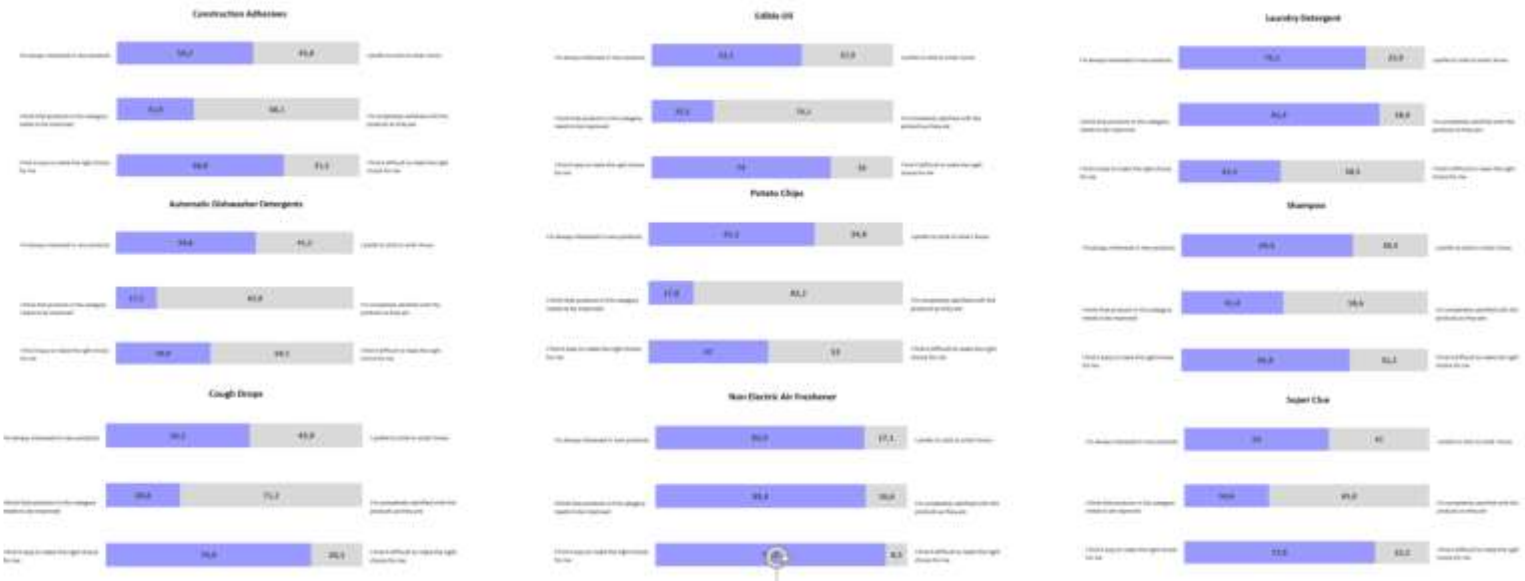
For “Role of Innovation” (RoI) the following three differentials are used:

“I’m always interested in new products”
vs
“I prefer to stick with what I know”

“I think products in this category need to be improved”
vs
“I’m completely satisfied with the products as they are”

“I find it easy to make the right choice for me”
vs
“I find it difficult to make the right choice for me”

With these differentials, we can derive insights about the opportunities for new products in the different categories.



“Satisfaction with current products” ranges from 16.6% (“NECA”) to 82.2% (“Potato Chips”), representing a large difference in suppliers’ opportunity to develop new products. Another example: 39% find it “easy to make the right choice” in the dishwasher detergent category, compared with 92% for “NECA.” This may indicate the need for differentiation, such as by developing and clearly communicating specific USPs for different products.

INITIAL CONCLUSIONS

Results from the nine "Behavioral Calibration Questions" indicate that they have a potential to differentiate between respondents' buying habits. Regarding our hypothesis, behavior during the most recent shopping trip (within a category) influences the answering behavior in the choice exercise. Considering this, these questions should help respondents to recall their decisions during their last shopping trip more effectively; therefore, responses to the following conjoint tasks should be much easier and clearer to them. Bivariate analysis of the "Behavioral Calibration Questions" suggests that our hypothesis may be correct and that it is worthwhile to invest the additional interview time to improve the answering behavior of respondents on the choice task.

ENHANCE CONJOINT

To explore how the calibration questions enhance the conjoint exercise that follow, we consider three different mechanisms:

- Simply asking the questions helps respondents to recall their most recent shopping trip, which results in more reliable answers.
- Using these questions as covariates improves the Bayesian estimation of the part-worth utilities and results in better "Share of Choice" estimations, better hit rates, and less error.
- The three roles may provide further insight when using them as segmentation variables in the choice simulator.

All nine empirical studies were analyzed using the same settings to avoid methodological bias. We used Sawtooth Software CBC/HB (190,000 burn-in-draws, write out 1,000 draws by using every tenth draw). For SoC simulation, we used the average over these 1,000 draws, as well as 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 three different estimations for the sample split cells with "Behavioral Calibration Questions":

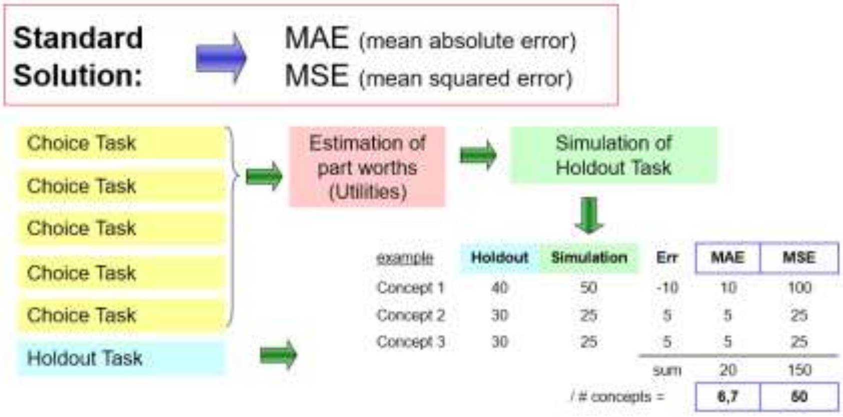
- Standard HB estimation,
- HB with the nine binary questions as covariates, and
- Ensemble of nine estimation runs with one of the questions used as a covariate in each run.

One of the great achievements of machine learning is certainly the use of ensembles. An ensemble approach generates multiple diverse models, include HB estimations with different covariates as in this study. First, we can make predictions with each of the specific HB models individually. Due to the different covariates, these models are diverse in the sense that each provides different predictions and has its own unique strengths and weaknesses. For the ensemble approach, we take the nine different models and blend the SoC predictions to reduce bias from the individual models, thereby generating more robust and accurate predictions.

SHORT REMINDER: HOW WE MEASURE THE VALIDITY OF CONJOINT STUDIES

Before describing the results of the different approaches, we would like to review how the measures of validity are computed.

The standard approach is to use one predefined choice task not used for the estimation process as a "holdout task." This task is then simulated and the MAE (mean absolute error) or the MSE (mean square error) for the whole sample is calculated taking into account the number of concepts in the task.



Standard Solution: MAE (mean absolute error) MSE (mean squared error)

Choice Task

Choice Task

Choice Task

Choice Task

Choice Task

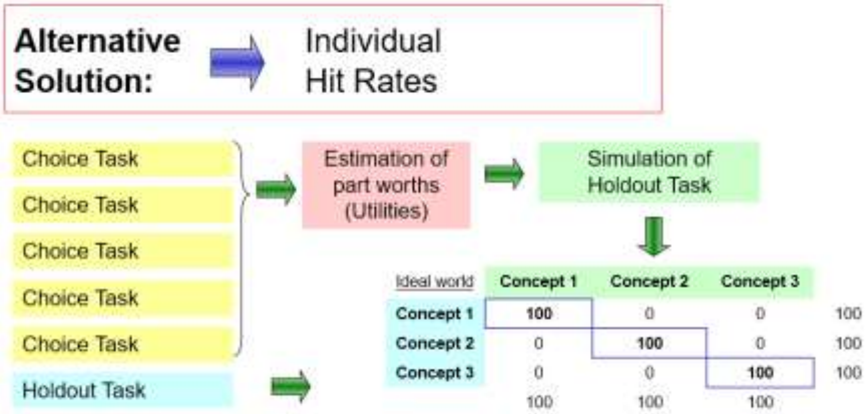
Holdout Task

Estimation of part worths (Utilities)

Simulation of Holdout Task

example	Holdout	Simulation	Err	MAE	MSE
Concept 1	40	50	-10	10	100
Concept 2	30	25	5	5	25
Concept 3	30	25	5	5	25
			sum	20	150
		/ # concepts =		6,7	50

The alternative approach is to individually simulate the "holdout task" for each respondent and match it with his or her actual answer to this task during the interview.



Alternative Solution: Individual Hit Rates

Choice Task

Choice Task

Choice Task

Choice Task

Choice Task

Holdout Task

Estimation of part worths (Utilities)

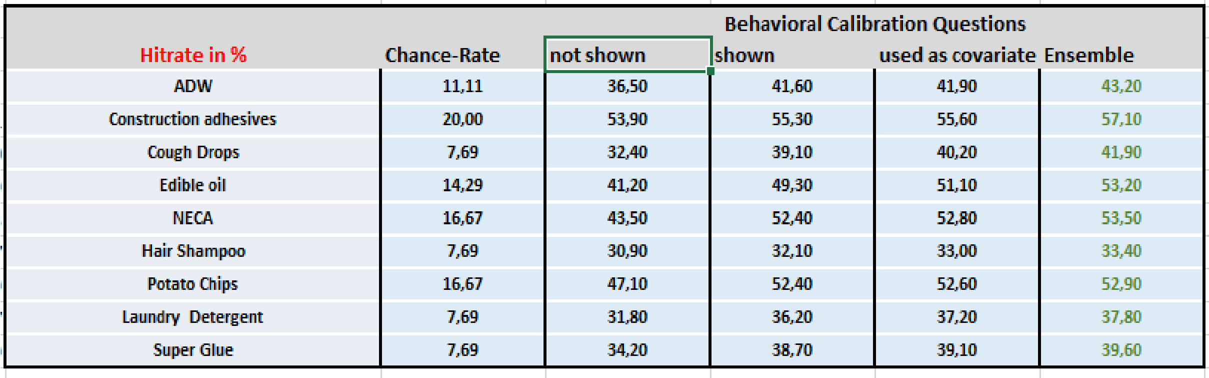
Simulation of Holdout Task

Ideal world	Concept 1	Concept 2	Concept 3	
Concept 1	100	0	0	100
Concept 2	0	100	0	100
Concept 3	0	0	100	100
	100	100	100	

If real market data (e.g., market shares) are available, the root mean squared error (RMSE) is used.

EMPIRICAL RESULTS

The nine studies analyzed confirm our hypotheses: the "Behavioral Calibration Questions" are effective, and hit rates can be significantly increased by asking these questions up front, even if they are not used in the HB estimation. Using the "Behavioral Calibration Questions" as covariates in the HB estimation further improves hit rates. Finally, an ensemble of part-worth utilities from nine estimations based on one calibration question as covariate in each run results in slightly higher hit rates compared with a single HB run that use all nine questions as covariates. Within a category, the more specific the three different roles of our behavioral calibration questions are, the more the hit rates can be improved by using this additional information in the estimation. For example, in the "NECA" category, where numerous consumers compare prices and search for new innovative products, the hit rate could be improved from 43.5% to 53.5%.

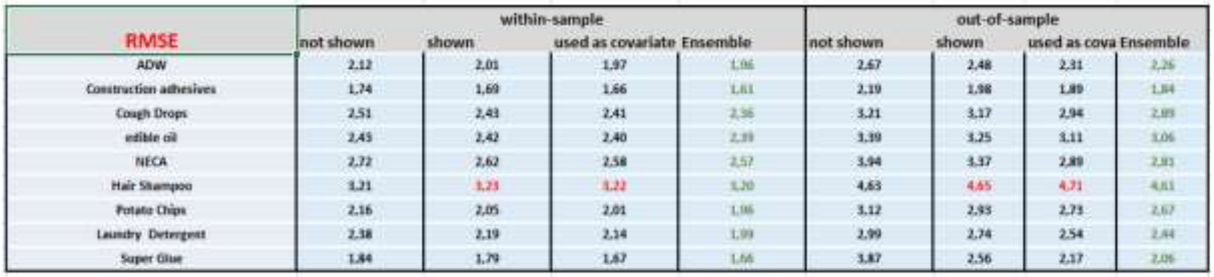


		Behavioral Calibration Questions			
Hitrate in %	Chance-Rate	not shown	shown	used as covariate	Ensemble
ADW	11,11	36,50	41,60	41,90	43,20
Construction adhesives	20,00	53,90	55,30	55,60	57,10
Cough Drops	7,69	32,40	39,10	40,20	41,90
Edible oil	14,29	41,20	49,30	51,10	53,20
NECA	16,67	43,50	52,40	52,80	53,50
Hair Shampoo	7,69	30,90	32,10	33,00	33,40
Potato Chips	16,67	47,10	52,40	52,60	52,90
Laundry Detergent	7,69	31,80	36,20	37,20	37,80
Super Glue	7,69	34,20	38,70	39,10	39,60

Out-of-sample calculations were done by splitting the samples into estimation and validation samples (80% and 20%, respectively).

Only the "super glue" study design consisted of four sample splits, such that the possibility of using validation samples and estimation samples was built into the design. In this study, the separate validation samples each had 250 respondents answering or not answering the Behavioral Calibration Questions, whereas the estimation samples had 500 respondents each.

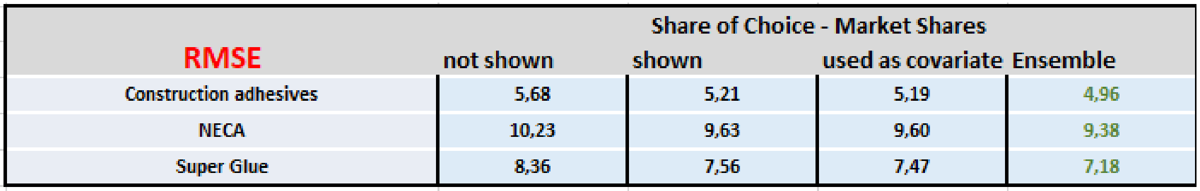
As we could not calculate part-worth estimates for out-of-sample tests, and therefore could not simulate preference shares in the traditional way, we used logCounts (described in Johnson, Orme, Pinnell 2006). The conclusion is roughly the same as that for the above-mentioned hit rates: almost all studies have better RMSE values when "Behavioral Calibration Questions" were implemented. Only the "Shampoo" study seemed to not benefit from use of the "Behavioral Calibration Questions," but the framing did not harm the results.



	within-sample				out-of-sample			
RMSE	not shown	shown	used as covariate	Ensemble	not shown	shown	used as cova	Ensemble
ADW	2,12	2,01	1,97	1,96	2,67	2,48	2,31	2,26
Construction adhesives	1,74	1,69	1,66	1,61	2,19	1,98	1,89	1,84
Cough Drops	2,51	2,43	2,41	2,36	3,21	3,17	2,94	2,89
edible oil	2,45	2,42	2,40	2,39	3,39	3,25	3,11	3,06
NECA	2,72	2,62	2,58	2,57	3,94	3,37	2,89	2,81
Hair Shampoo	3,21	3,23	3,22	3,20	4,63	4,65	4,71	4,61
Potato Chips	2,16	2,05	2,01	1,96	3,12	2,93	2,73	2,67
Laundry Detergent	2,38	2,19	2,14	1,99	2,99	2,74	2,54	2,44
Super Glue	1,84	1,79	1,67	1,66	3,87	2,56	2,17	2,06

Only two of our nine studies have reliable, "real" market shares information and can therefore be compared against them. In the "super glue" case, we estimated separate models for the validation and estimation splits and compared them with the RMSE measure.

The results support the same deductions as the above comparisons: simply asking the “Behavioral Calibration Questions” improves the predictions. The inclusion of these questions as covariates or in an ensemble approach further improves the “Share of Choice” simulations.

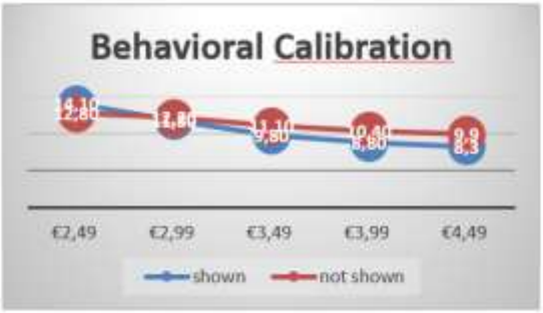


	Share of Choice - Market Shares			
RMSE	not shown	shown	used as covariate	Ensemble
Construction adhesives	5,68	5,21	5,19	4,96
NECA	10,23	9,63	9,60	9,38
Super Glue	8,36	7,56	7,47	7,18

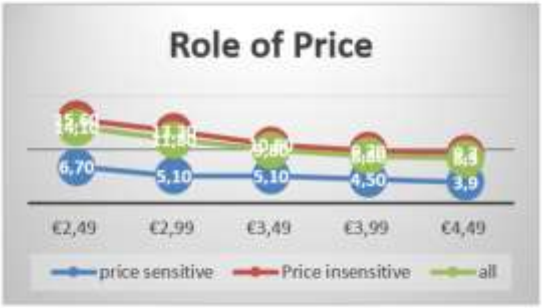
USE BEHAVIORAL CALIBRATION AS SEGMENTATION

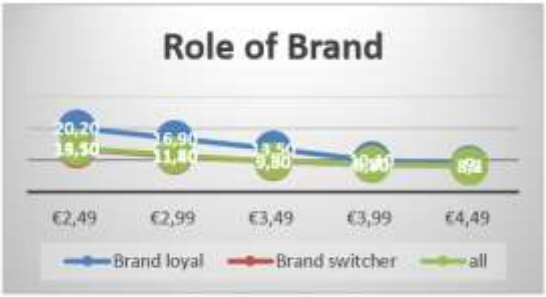
Our third approach in using the “Behavioral Calibration Questions” is based on the three “Roles.” For each, we calculated a filter variable based on the three questions to derive specific “Share of Choice” values for the splits.

The following example shows different price elasticities for one SKU in the “edible oil” study. Again, it is clear that asking the “Behavioral Calibration Questions” results in different elasticities:

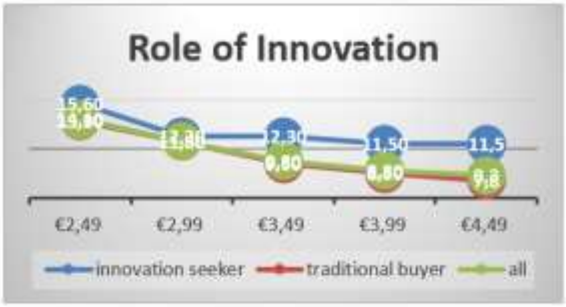


The different elasticities correspond with our expectations regarding the role of price, in that price-sensitive buyers with brand-switching behavior (i.e., respondents who switch to a different brand when price increases) have higher elasticities:

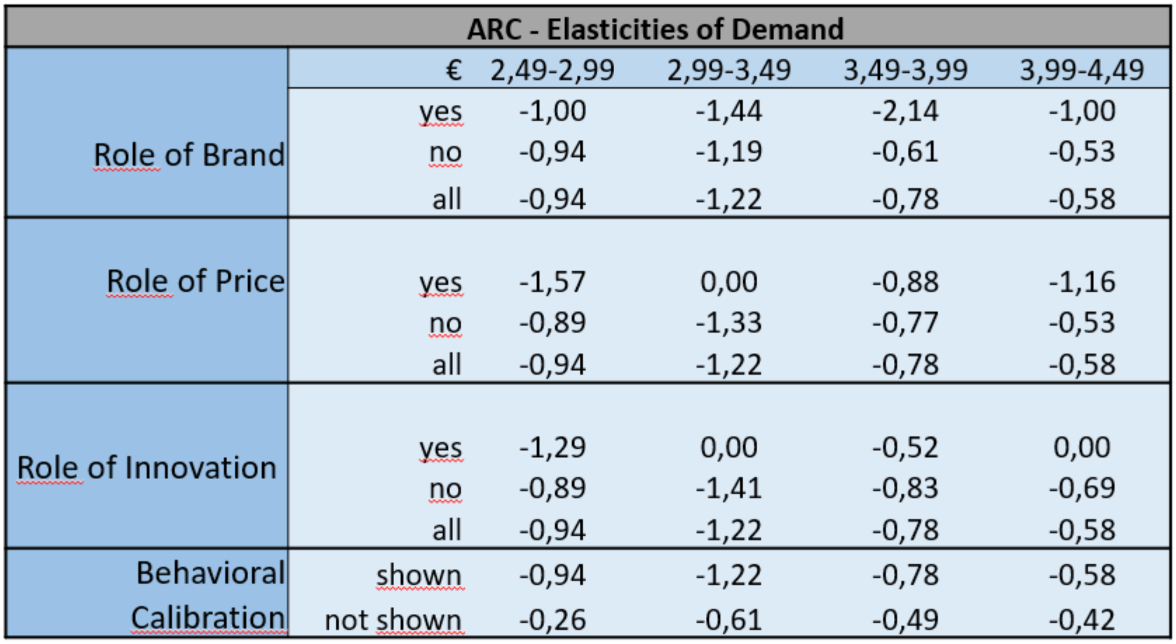




Innovation seekers are less price sensitive. Simply exposing respondents to the "Behavioral Calibration Questions" results in different elasticities ("edible oil" study):



For a more detailed inspection of these effects, we calculated the arc-elasticities of demand for the different segments. The differences between the segments provide detailed insights into the influence of consumer behavior on price and can be leveraged for more insightful recommendations to clients.



ARC - Elasticities of Demand					
	€	2,49-2,99	2,99-3,49	3,49-3,99	3,99-4,49
Role of Brand	yes	-1,00	-1,44	-2,14	-1,00
	no	-0,94	-1,19	-0,61	-0,53
	all	-0,94	-1,22	-0,78	-0,58
Role of Price	yes	-1,57	0,00	-0,88	-1,16
	no	-0,89	-1,33	-0,77	-0,53
	all	-0,94	-1,22	-0,78	-0,58
Role of Innovation	yes	-1,29	0,00	-0,52	0,00
	no	-0,89	-1,41	-0,83	-0,69
	all	-0,94	-1,22	-0,78	-0,58
Behavioral Calibration	shown	-0,94	-1,22	-0,78	-0,58
	not shown	-0,26	-0,61	-0,49	-0,42

FINDINGS

Based on our nine empirical studies, we can conclude that the "Behavioral Calibration Questions" represent a useful extension to DCM exercises. Our findings suggest that all three hypotheses may be verified. The "Behavioral Calibration Questions" help the respondents to recall their most recent shopping trip in a particular category and thereby positively influence answering behavior in the ensuing conjoint model. The data-generation process comes closer to representing a real shopping trip. Using the questions as covariates can also help improve the estimation results, rendering them more meaningful for simulations. The use of nine different estimations based on the "Behavioral Calibration Questions" in an ensemble approach slightly improves the results and always performs slightly better than a single estimation with nine covariates. Due to the modest improvement, one should decide if the additional effort required by this approach is justified. Using the "Behavioral Calibration Questions" as filter variables provides more detailed insight into the data structure and helps to improve recommendations for clients.

Consequently, it seems that further investing in these additional questions is worthwhile to improve our conjoint models.

FUTURE RESEARCH

The nine "Behavioral Calibration Questions" are a good starting point for further developments. To take advantage of such a framing exercise, the "Behavioral Calibration Questions" could be extended to include more than the three roles. For instance, three additional semantic differentials about the importance of features could be added to generate a fourth role. More specific wording for different categories should also be developed and validated.

From a more methodological point of view, a next step could be the use of the questions as an input for a Bayesian variable selection model to improve part-worth estimates.

Benchmarks should be established by building a database of Category Behavioral Calibration results to position tested concepts in commercial studies.

Peter Kurz

Stefan Binner

References

Allenby, G.M.; Rossi, P.E. (2006): Hierarchical Bayes Models, in: Grover, R.; Vriens, M. (Eds.): *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*, 418–440, SAGE Publications Inc., Thousand Oaks.

Binner, S. (2006): Do Individual Hit Rates Matter at All? (Design & Innovations Conference München 2006).

Hein, M.; Kurz, P.; Steiner, W. (2013): Limits for Parameter Estimation in Choice-Based Conjoint Analysis: A Simulation Study, European Conference on Data Analysis 2013.

Hein, M., Kurz, P., Steiner, W. (2019): Analyzing the Capabilities of the HB Logit Model for Choice-Based Conjoint Analysis: A Simulation Study. In: Journal of Business Economics (forthcoming).

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

Kurz, P; Binner, S. (2011): Added Value through Covariates in HB Modeling?, Proceedings of the 2011 Sawtooth Software Conference.

Liakhovitski, D.; Shmulyian, F. (2011): Covariates in Discrete Choice Models: Are They Worth the Trouble? ART Form Presentation.

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

Sentis, K. and Li, L. (2001): One Size Fits All or Custom Tailored: Which HB Fits Better? Proceedings of the 2001 Sawtooth Software Conference.

Sentis, K.; Geller, V. (2011): The Impact of Covariates on HB Estimates, Proceedings of the 2011 Sawtooth Software Conference.

ENHANCEMENT OF VAN WESTENDORP PRICE MODEL VIA NEWER STATISTICAL APPROACHES

MING SHAN
KYNETEC

BACKGROUND AND OBJECTIVES

Over four decades since its original introduction, the van Westendorp Price Sensitivity Meter (Van Westendorp 1976, abbr. as **PSM**) remains as a frequent method choice in survey-based pricing research. The four questions required can be easily collected via different survey modes. The analysis and output interpretations are relatively simple and straightforward. A PSM extension by Newton, Miller, and Smith (Miller et al. 1993, abbr. as **NMS**) by adding two additional purchase intention questions further allows a price-demand curve to be plotted.

While there is no shortage of critiques to the method, the fact of its widespread usage means any enhancement or even advice on potential pitfalls would not be a total waste of effort. After all, the pricing decision is critical in market research and an inferior recommendation could result in substantial financial loss without being even noticed. While efforts aimed at applying more stringent statistical approaches have been made (e.g., Lipovetsky 2006), the dominant usage seems to have been following its original form.

This paper shares some work on exploring and building statistical models using PSM and NMS data. Statistical options considered include survival analysis, multi-level modeling and Bayesian inference. We also raise questions about some shortcomings of the existing approaches and offer suggestions to improve.

R is an open-source platform for statistics and graphs (R Core Team 2021) and data sciences. To facilitate easy usage and promote best practices, an R package will be developed to share.

PSM AND NMS—A VERY BRIEF SUMMARY

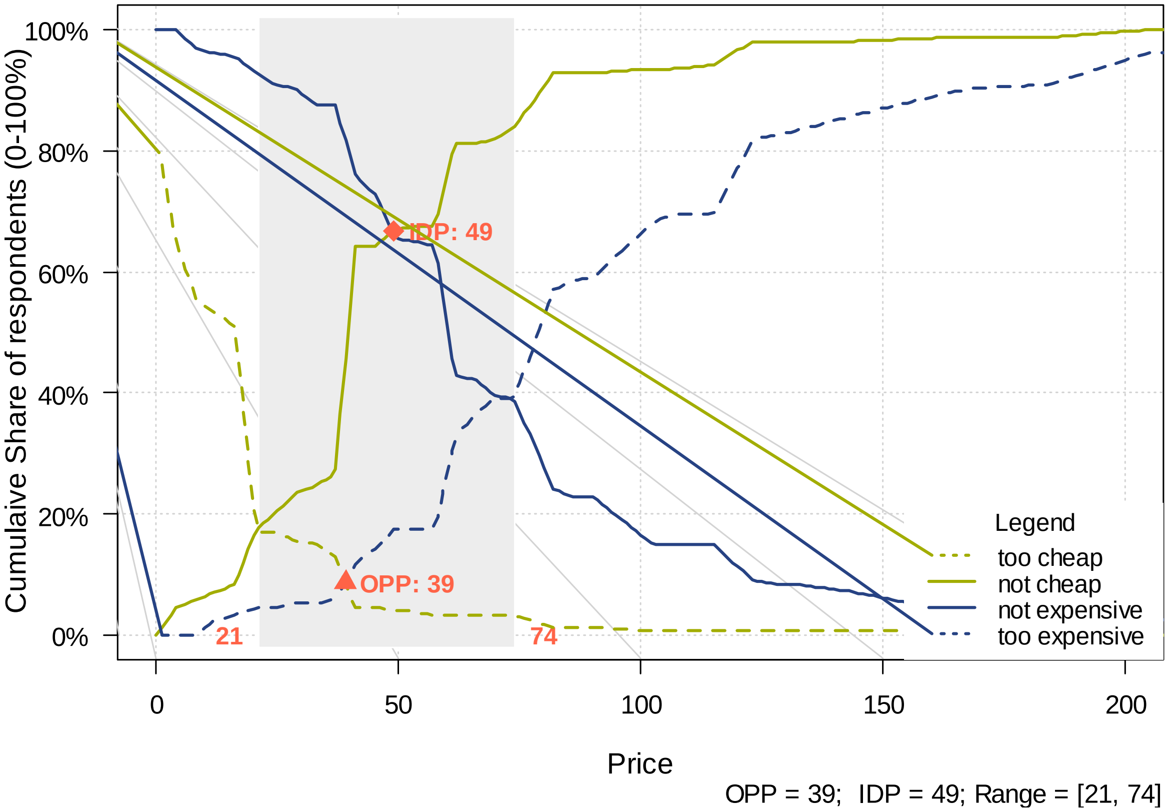
PSM uses 4 questions to construct a price scale:

At which price are you beginning to experience Product X as:

1. **too cheap**—so that you say "at this price the quality cannot be good"
2. **cheap**
3. **expensive**
4. **too expensive**—so that you would never consider buying

Figure 1 is a typical PSM output. Van Westendorp's idea was to construct 4 cumulative distribution curves separately from the 4 reported prices and then use the different intersecting points to define those so-called *optimal price, indifference price,* and *marginal cheap* and *marginal expensive* prices. A minor point to note here is that we follow Van Westendorp's original proposal by reversing the distributions on "cheap" and "expensive" to show "not-cheap" and "not-expensive" curves instead.

Figure 1: Traditional PSM

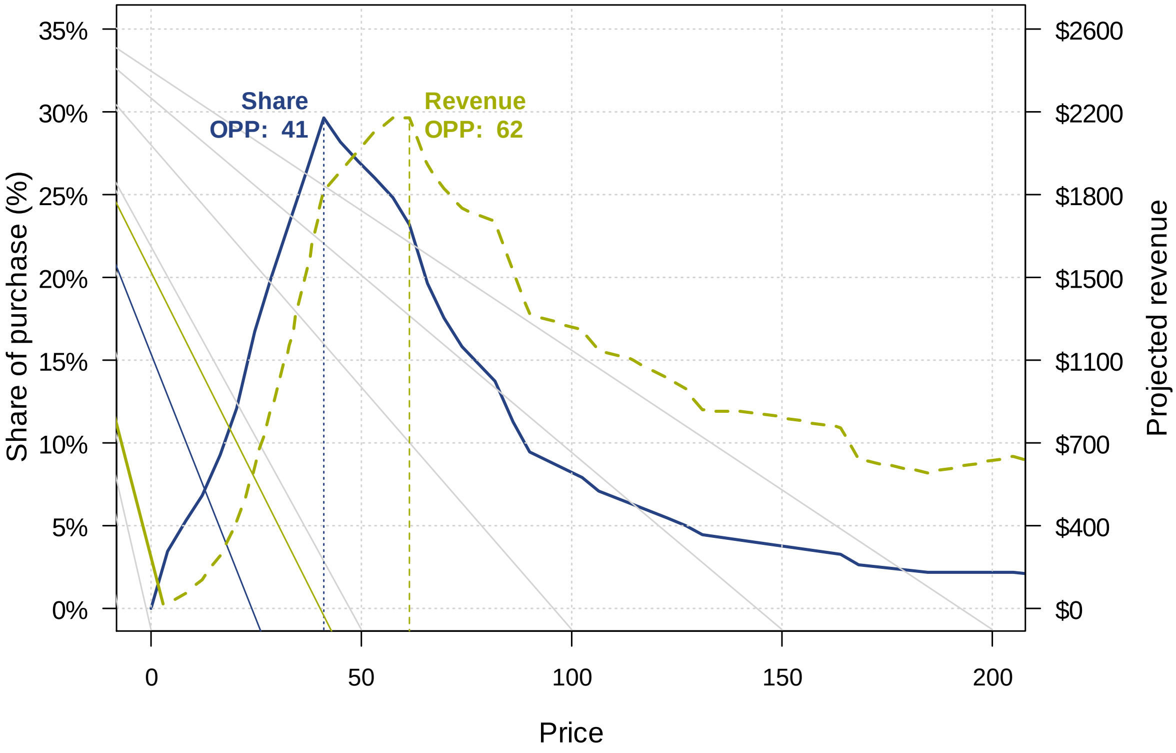


Once a respondent has named the 4 prices following PSM, NMS asks each respondent two additional questions on purchase intention:

> *How likely would you be to buy Product X at **inexpensive (or cheap)** price and **expensive** price:*
>
> *definitely would buy*
> *probably would buy*
> *might or might not buy*
> *probably would not buy*
> *definitely would not buy*

Figure 2 illustrates the result from NMS. In combination with the 4 original PSM questions, the two purchase intention questions help create a curve on the share of intended purchase and another curve on the revenue as the price increases. These curves allow one to look for their peaks to identify the corresponding optimal prices. To save space, we show the two curves in a single chart.

Figure 2: Newton-Miller-Smith Extension



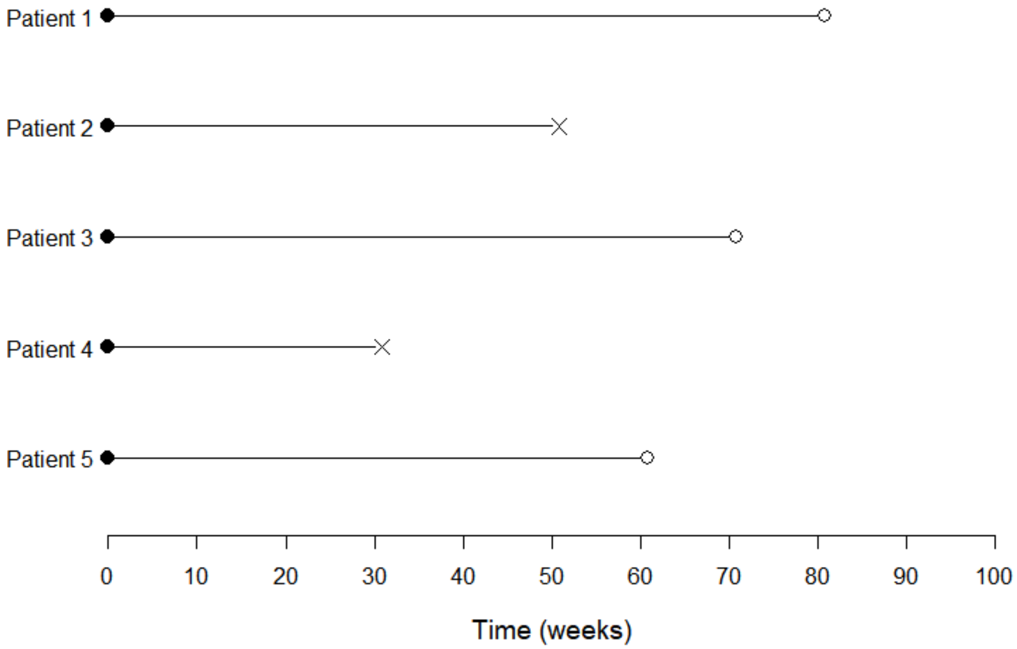
Over the years, there have been questions about the lack of theoretical support of PSM. The result is static and additional statistical measurements like confidence intervals are absent. It seems that these intersection points could be heavily driven by partial data. These four independent distributions mean PSM looks at data **across** but not **within** respondents, leaving a lot of information unused. One can also question the stability of its result and if there is any concern of bias.

For NMS, we are going to raise a major question shortly regarding the purchase intention assumption and show it generates an overstated optimal price by a large amount, thus we recommend a correction.

Survival Analysis

Survival analysis is applied in many fields such as medicine, biology, public health, epidemiology, engineering, demography, and economics (Klein 2003). In social sciences such as demography and economics, the method is also called event-history analysis. The key concerning variables include *duration to event* which we use *X* to denote, its *censoring status* plus some other *covariates* such as the previous medical conditions or treatment levels of patients or comparison groups in a controlled experiment. Censoring occurs if the final survival status cannot be fully observed or followed through, for example when the event has yet to happen before the termination of a study. Figure 3 is an example on the survival status among 5 patients. Patient #2 and #4 could be those dropped out of a study for different reasons. Fortunately, censoring is not a concern for our problem, so we can pretty much ignore it. The key in our application is to treat **price as the failure time X.**

Figure 3: An Illustration of Survival Status of 5 Patients



Just very briefly, the distribution of *X* is characterized by four different functions: survival function, hazard rate (function), probability function and mean residual life at X. All these key functions which summarize the behaviors of the survival data are interlinked. Knowing one of them, the other three are determined. The survival function has a range between 1 and 0 for percentage. One of the most concerning functions is called the hazard function, which specifies the conditional failure rate at any given point of time.

There are different and well-established approaches for survival analysis/modeling. The Kaplan-Meier curve is a classic non-parametric option. The well-known Cox proportional hazards regression analysis is a semi-parametric approach. With explicit assumption on distribution and time, varying covariates one can also take a parametric approach. Lastly, machine learning using survival tree/survival random forests is also possible.

Van Westerndorp via Survival Analysis

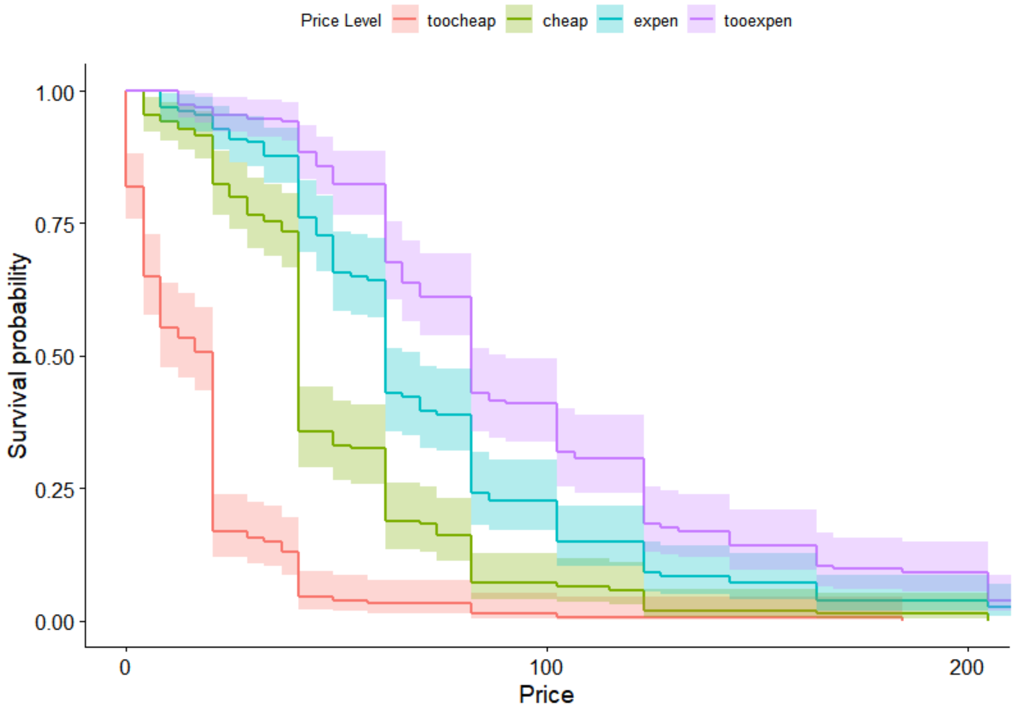
To fit PSM data into survival framework, we can view price X (X>0) as "time." Specifically, we have 4 prices from low to high as X1 to X4:

- X1 = Too cheap, X2 = Cheap, X3 = Expensive, X4 = Too expensive
- Also satisfy: X1 < X2 < X3 < X4

We treat each named price point from PSM as the event or failure since it is when we expect to lose a customer because the price is no longer acceptable, or as survival since we can win/keep the customer below that price as it is still satisfactory.

Traditional PSM is pretty much the result of a summary of the four separate survival curves as illustrated in Figure 4. One way to think of it is to estimate the "average" or "median," or some sort of baseline survival curve expected to fall between the green and blue lines in the demo plot. For example, we can fit a Cox proportional hazard model by first stacking the data and then treating the 4 prices as a categorical covariate. Or we could simply average X2 and X3 of each respondent as "time" variable to construct a survival curve. Once a survival curve is constructed, we can then calculate the revenue curve ($= S(X) \times X$) and locate the price producing the highest revenue. Although not shown here, other functions of price including price elasticity can be derived.

Figure 4: Kaplan-Meier Curves for 4 PSM Prices



Newton-Miller-Smith Extension

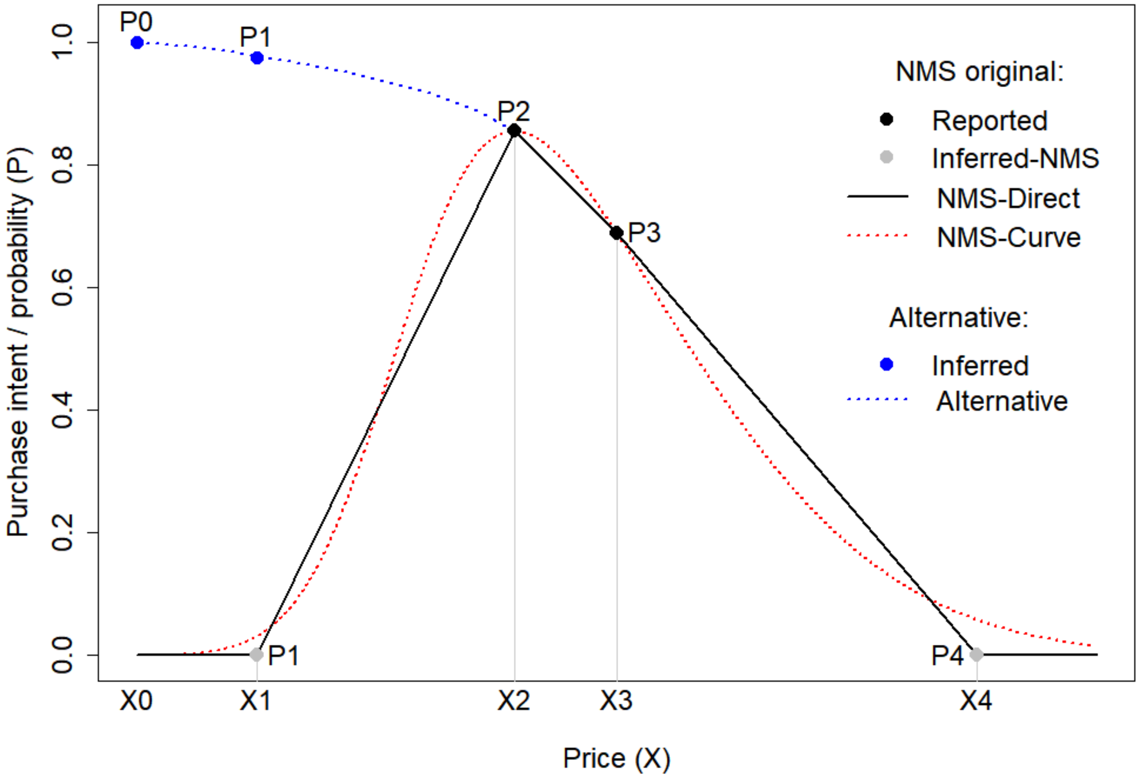
Questioning the Original Assumptions

In NMS, each respondent gives two purchase intention values, we call them P2 and P3, both falling between 1 and 0 if we view or convert them into a probability. According to the original paper, although not directly asked, P1 and P4, which are the purchase intention at "too cheap" and "too expensive price," are set to 0. For each respondent, this results in a purchase intention curve shown in the chart in Figure 5 by the solid black line. A smoothed curve illustrated by the red dotted line can also be fitted.

PSM is often used to test some new and likely superior products. One can argue if we should expect a higher purchase likelihood of P1 when the price is cheaper, all other things equal. If we accept this argument, it means that by heavily suppressing the purchase likelihood at the lower price zone NMS understates the market potential at lower price space and inflates the final recommended price.

This leads to our proposal to modify the original NMS. **We set purchase probability P0 when price is equal to 0 to 1 instead of 0 since the product is completely free**. We then interpolate P1 using P0 we just set plus P2 reported by the respondent. Same as NMS, P4 is set to 0 or near 0. This gives us a few data points to fit an individual demand curve for each respondent. The dotted line in blue is the proposed alternative curve. With this assumption, each individual respondent now has a waterfall shaped survival curve.

Figure 5: Purchase Intent of a Respondent—Alternative

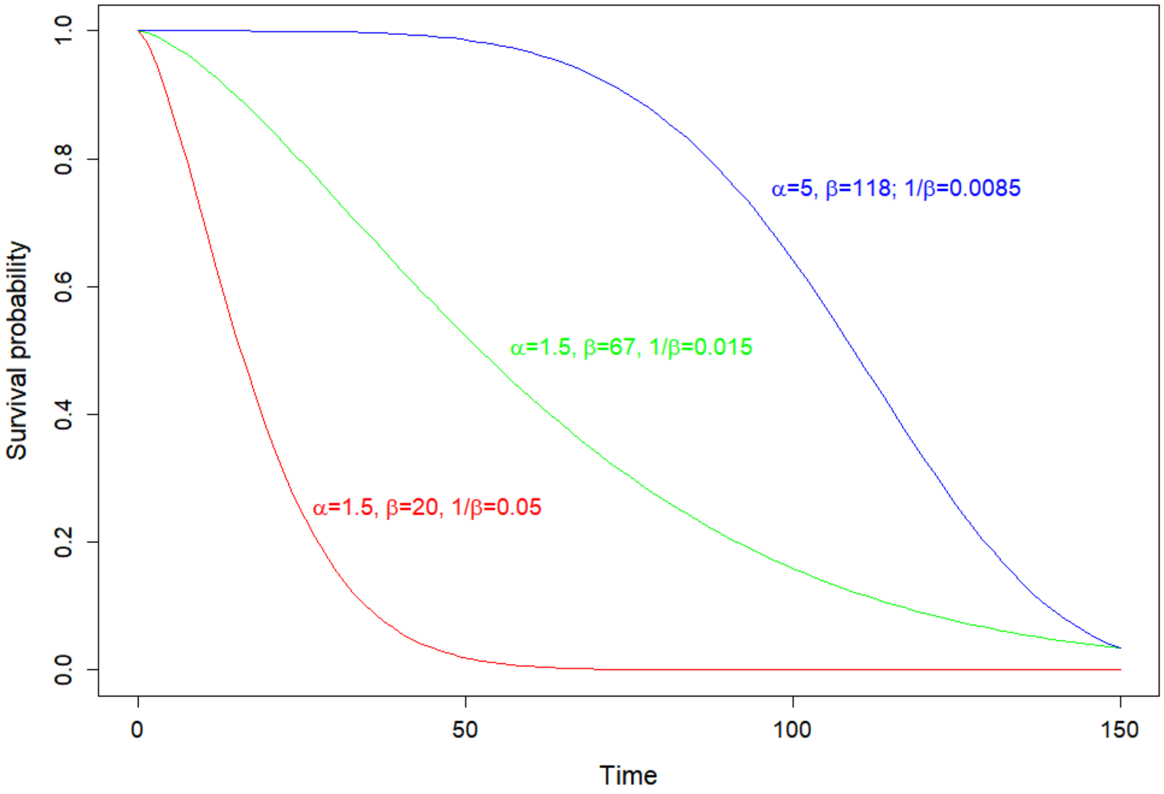


Need of Modeling

Besides arguing that we should amend the assumption, we also would like to seek improvement from a modeling perspective. The original NMS paper appears to be non-parametric by aggregating some zigzagged lines across individuals. For example, there is no mention on how one should determine the exact price points when P2 and P3 are the same, implying a range of optimal prices instead of a single unique one for a respondent. But how to pick a single price will have implication on the value of the final recommended price.

We propose a parametric model on NMS data using a Weibull distribution which is well applied in survival analysis, to express the waterfall shaped curves. We choose its form with 2 parameters α and β, both positive. Variable x which is greater than or equal to 0 is the time to event. Also, α and β are parameters on "shape" and "scale" respectively. Figure 6 shows 3 different survival curves corresponding to different α and β values. This shows each customer's purchase intention curve can be characterized with only 2 values. Since we establish it in mathematical forms, if we wish, all other survival related features can also be easily obtained through mathematical expressions.

Figure 6: Weibull Survival Function with Different Shape and Scale Parameters



A Weibull survival function can be expressed as the following, where α > 0 is the shape parameter and $\beta > 0$ is the scale parameter:

$$y = e^{\left(-\frac{x}{\beta}\right)^{\alpha}} \quad (1)$$

Or $$-\ln y = \left(\frac{x}{\beta}\right)^{\alpha} \quad (2)$$

Or $$\ln(-\ln y) = \alpha \ln(\frac{x}{\beta}) \quad (3)$$

Or $$\ln(-\ln y) = \alpha \ln(x) - \alpha \ln(\beta) \quad (4)$$

Or $$y' = a\, x' - b \quad (5)$$

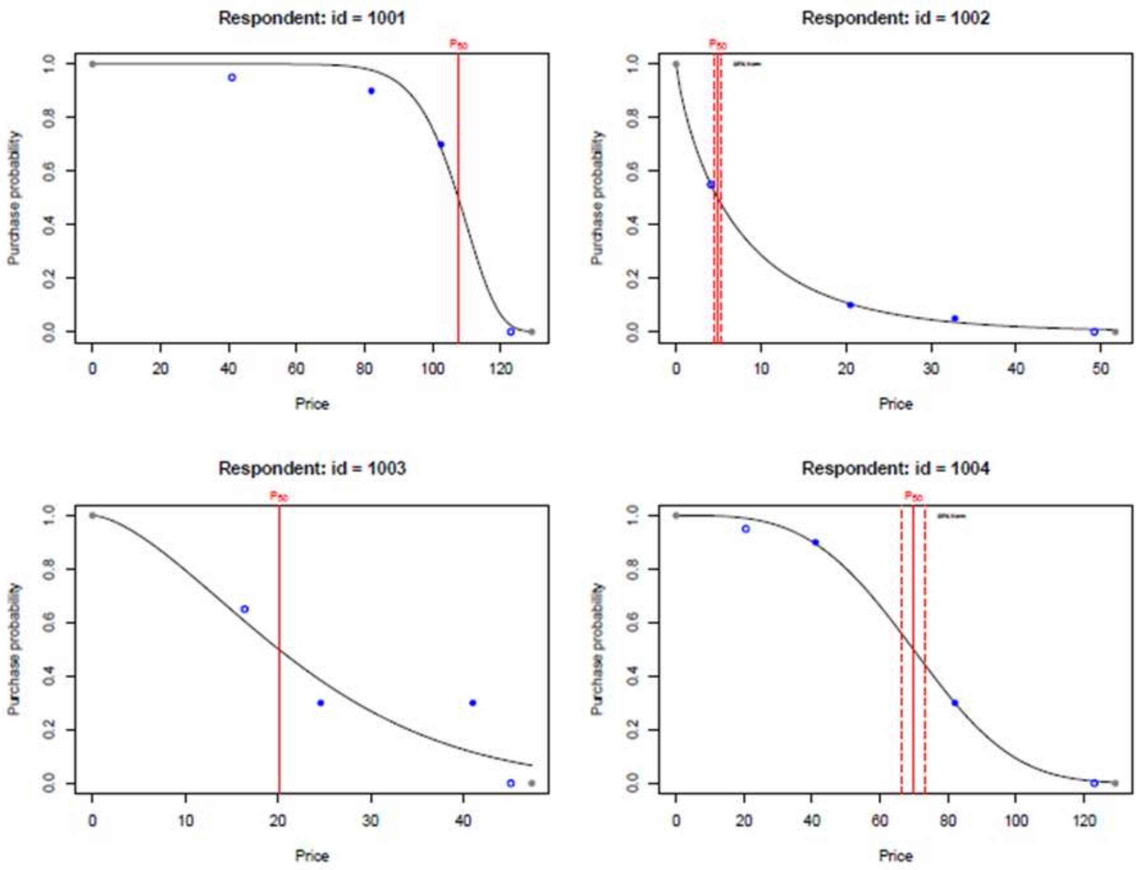
To fit a survival curve with Weibull function on data points shown in Figure 5, we can directly estimate Equation 1 using a non-linear mixed effects model (Duursma and Choat 2017, Pinheiro and Bates 2000). After a couple of *ln* transformations, which leads to Equation 5 as shown, we can also fit a simple linear regression on the transformed data and backtrack to the final α and β values. Fitting a single multilevel/hierarchical (or mixed-effects) model on the whole study sample, which leverages pooling/information borrowing across respondents to obtain both group-level and individual estimates, has many advantages over fitting a separate model for each respondent independently based on only a few data points (Gelman and Hill 2006). We use the R *nlme* package to fit a nonlinear mixed-effects model (Pinheiro et al. 2021).

One might speculate that some alternative model forms are possible. Besides some of the attractive features of the Weibull model for interpretation, Weibull models seem to fit the type of data addressed here better than logistic models based on some comparisons made but not reported here.

Illustration of Potential Applications

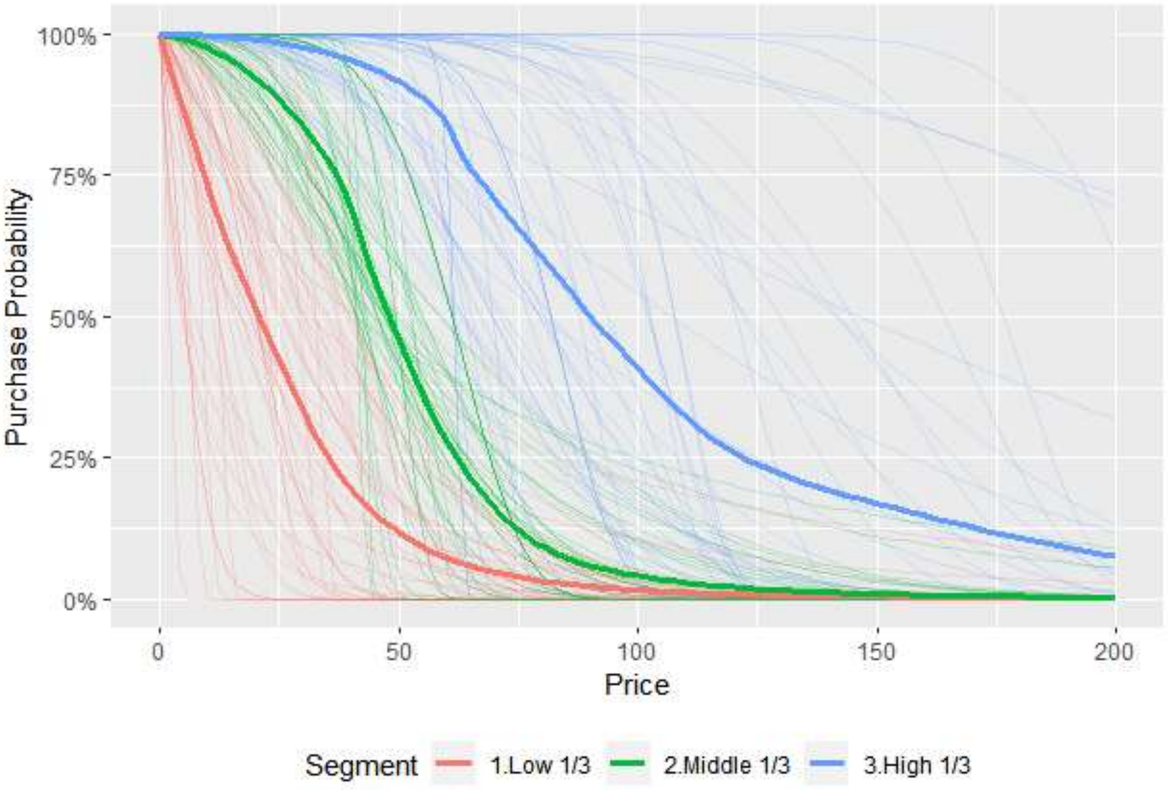
Figure 7 shows what the curves look like for 4 different individuals. Once we are happy with the individual curves, the total results are just a matter of aggregation of these individuals to produce different types of results, like a probability curve or a revenue curve.

Figure 7: Examples of Individual Purchase Probability Curves



With only 2 parameters, the Weibull model is quite parsimonious and can concisely characterize the pricing preference pattern of a respondent. For example, we can create segmentation by running a simple clustering analysis on the 2 parameters. Figure 8 shows an example when the total sample is divided into 3 equal segments just on the scale parameter alone. Each thin line is the purchase intention curve of a single respondent. The 3 thick lines are the segment centers. Of course, such detailed information at the respondent level supports other additional interpretation or analysis on pricing questions. For example, for a given price that a company intends to set, we can easily identify the percentage of the respondents exceeding the price threshold and then further understand those customers by linking to other respondent-level characteristics.

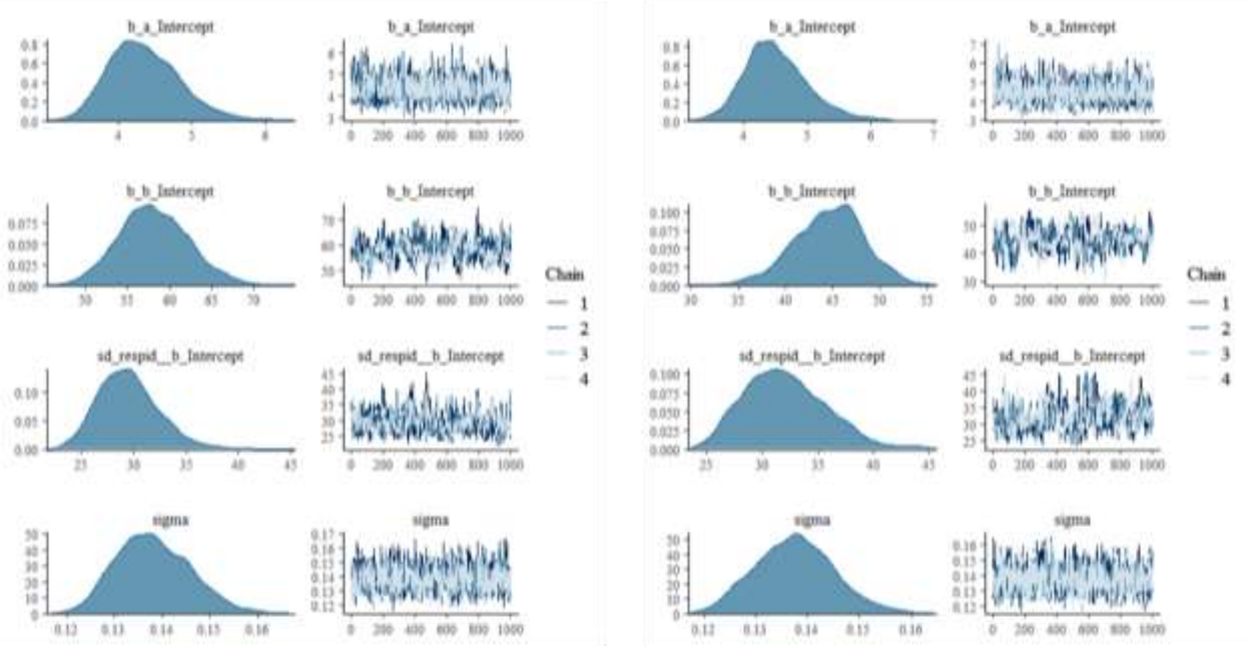
Figure 8: Segmenting Respondents on Their Individual Purchase Intention Curves



Bayesian Estimation and Inference

We also ran Bayesian estimation on the same model using *Stan* (Carpenter et al. 2017), a Probabilistic Programming Language for Bayesian estimation. R package *brms* (Bürkner 2017) makes it simple by allowing all model specifications to be made within R. Figure 9 compares the posterior distributions of model parameters between two different choices of priors. The left panel shows the results when flat or less informative priors are applied to let our results be mostly dominated by the data. The right panel shows when we set the prior on the scale parameter with a lower price plus a narrower distribution, allowing the prior to have a stronger influence on the posterior distribution. If we compare the two posterior distributions in the second row, which is about the scale (or the level) of price, we get an optimal price of 58 on the left and an optimal price of 47 to the right. What we intend to demonstrate here is that business hypothesis or knowledge about price can be injected in the form of prior to mix with survey data under the Bayesian framework, especially when we are less confident about the data, in situations like low sample size or having concerns about the data quality.

Figure 9: Posterior Distribution Comparison—Weaker vs. Stronger Priors



Using weaker priors

Using stronger priors

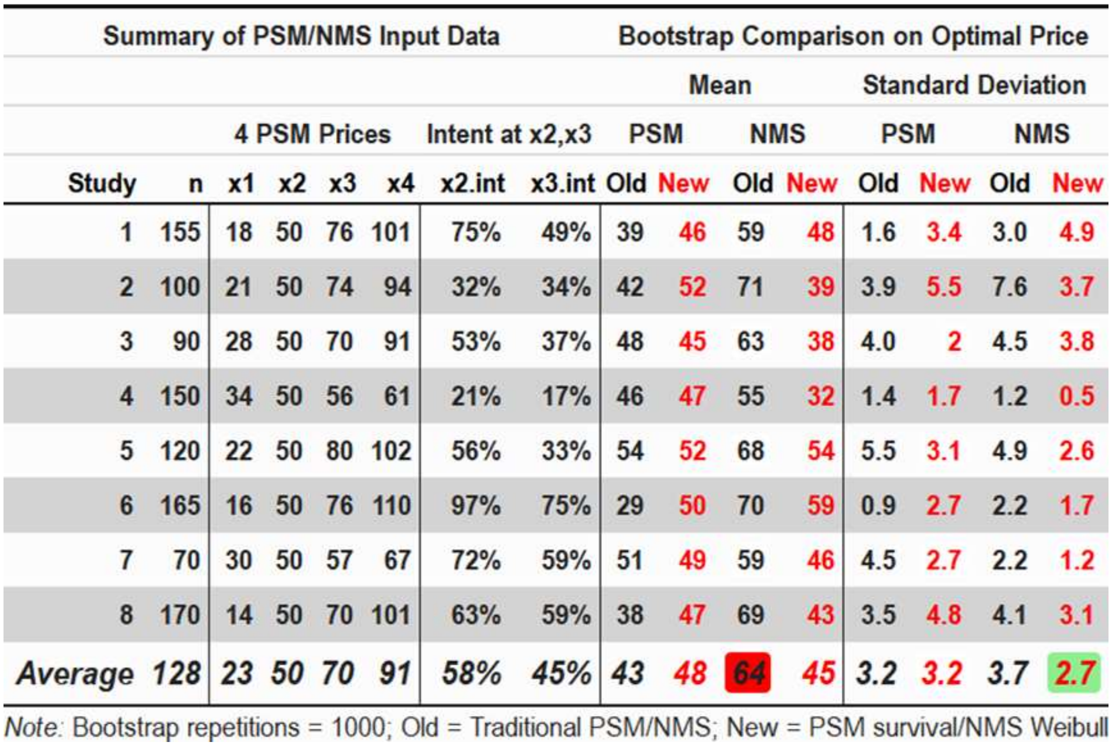
A few points are worth mentioning as a conclusion on the proposed improvement on NMS. All individual-level estimation can be easily aggregated into the total. A mixed-effects model allows information borrowing and data smoothing, balancing between individual heterogeneity and group-level coherence by tuning down the individual-level noise. Although not shown here, other detailed information such as price elasticity, respondents by percentiles can be easily obtained. Lastly, other purchase intention scales with range not limited between to 0 to 1 or respondent weighting can be implemented independent of the framework we are proposing.

EMPIRICAL DATA EVALUATION

Data and Measurements

In this section we compare the existing approaches on PSM and NMS with each of the corresponding modifications that we are proposing, *traditional PSM* vs. *PSM survival model* and *original NMS* vs. *NMS Weibull.* A total of 8 data sets were gathered, all including traditional 4 PSM questions plus two NMS intention questions measured by intended purchase shares. These are all international studies varying by different types of respondents and product categories. To aid comparison, all prices are rescaled by forcing the average of “cheap” price $X2$ equal to 50. The focus is on the optimal pricing point that produces the highest revenue and their standard deviation based on bootstrap with 1000 draws each. Table 1 summarizes the result.

Table 1: Comparison among 4 Methods on 8 Different Studies



Summary of PSM/NMS Input Data								Bootstrap Comparison on Optimal Price							
								Mean				Standard Deviation			
		4 PSM Prices				Intent at x2,x3		PSM		NMS		PSM		NMS	
Study	n	x1	x2	x3	x4	x2.int	x3.int	Old	New	Old	New	Old	New	Old	New
1	155	18	50	76	101	75%	49%	39	46	59	48	1.6	3.4	3.0	4.9
2	100	21	50	74	94	32%	34%	42	52	71	39	3.9	5.5	7.6	3.7
3	90	28	50	70	91	53%	37%	48	45	63	38	4.0	2	4.5	3.8
4	150	34	50	56	61	21%	17%	46	47	55	32	1.4	1.7	1.2	0.5
5	120	22	50	80	102	56%	33%	54	52	68	54	5.5	3.1	4.9	2.6
6	165	16	50	76	110	97%	75%	29	50	70	59	0.9	2.7	2.2	1.7
7	70	30	50	57	67	72%	59%	51	49	59	46	4.5	2.7	2.2	1.2
8	170	14	50	70	101	63%	59%	38	47	69	43	3.5	4.8	4.1	3.1
Average	***128***	***23***	***50***	***70***	***91***	***58%***	***45%***	***43***	***48***	***64***	***45***	***3.2***	***3.2***	***3.7***	***2.7***

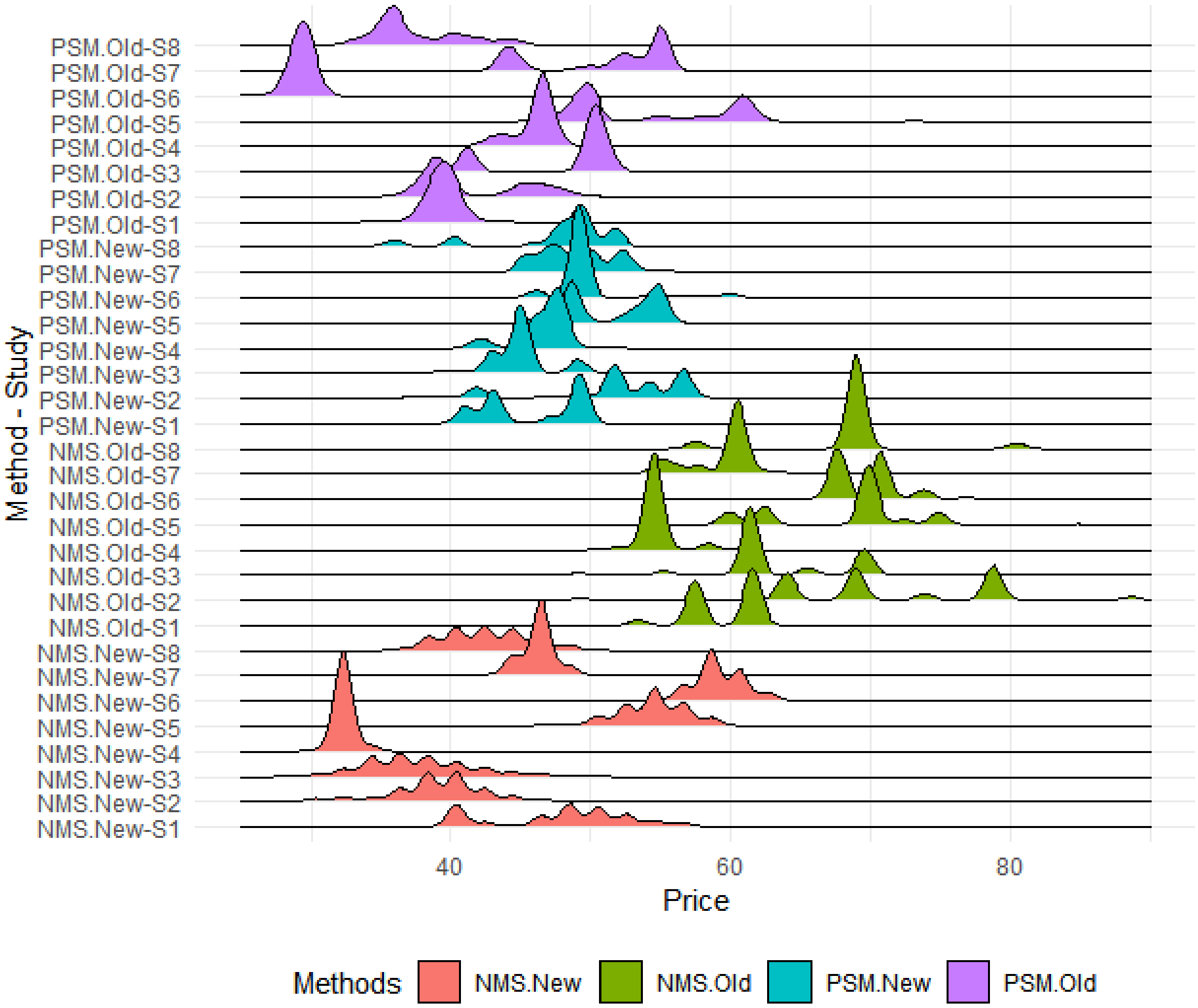
Note: Bootstrap repetitions = 1000; Old = Traditional PSM/NMS; New = PSM survival/NMS Weibull

The first observation goes to the highlighted number in red, which is the average optimal price of 64 from NMS. This is at least 1/3 higher than the next highest optimal price of 48 from the other 3 methods, quite a striking difference. NMS also has the largest standard deviation.

Among all 4 options, only traditional PSM is positively correlated with "too cheap" price X1 ($r = 0.67$), suggesting it might be over-influenced by that price. Also, the optimal price from traditional PSM has negative or no correlation with those from the other methods. Lastly, the results between the two old methods are negatively correlated ($r = -0.36$) while the results from the two new methods are positively correlated ($r = 0.43$), suggesting some convergence between the two new methods.

Figure 10 shows the detailed bootstrap distributions among the 4 methods on each data set. NMS in green stands out in positions to the right due to its high prices. Some of the multimodal distributions suggest a need for bootstrap check in general.

Figure 10: Bootstrap Distribution of Optimal Price (4 Methods by 8 Studies)



Implementation—R Package

A package in open-source R will be shared on GitHub (https://github.com/mingshan-mds) for free access. Upon favorable acceptance, package submission will be made to the formal R package repository (https://cran.r-project.org/web/packages). The plan is to offer at least the following features in the initial version:

- Implementation of the core methods/calculations proposed in this paper
- Simple and intuitive function calls
- Documentation and examples
- Detailed outputs and report-ready graphics
- Diagnostics such as bootstrap confidence intervals
- Sub-group analysis

To keep the package self-contained and easy to use without a need for advanced knowledge of R, we plan to leave the Bayesian estimation out of the package at least initially due to the requirement special software installation and longer running time. However, we still plan to share some more detailed descriptions and examples on Github for those who are interested in exploring further about the Bayesian application especially related to more intuitive prior setup.

Summary and Recommendations

Methodological

Survival analysis is proposed in this paper as the general foundation of PSM/NMS analysis for a couple of main reasons. First, the four PSM cumulative distribution curves are very similar to the classic survival curve. So is the proposed purchase intention curve in NMS except that it is at the individual level. Second, survival analysis offers a wide range of available theoretical guidance and calculation tools. Mixed-effects models focus on data within respondents instead of across respondents to allow more insight extraction, especially modeling individual price preference patterns. Bootstrapping and the use of more formal statistical models allow the outcomes to be expressed via distributions (akin to posterior distribution in Bayesian inference) for easy interpretation and expression of the level of uncertainty of the data and the model output. This in turn offers more prudent analytical input for the final pricing decision-making. We also hope some of the statistical models articulated here provide hints and motivations for further modeling expansions.

With the intention to improve the analysis under their original framework, the proposed method modification is built on the original survey questions from PSM and NMS. We could also consider in the opposite direction by adjusting how and what questions should be asked in the first place to capture the price perception of the respondents more precisely and also to better support the model building. For example, if the "too cheap" price is ambiguous to the respondents and since as the lowest price it is less supportive to the construction of individual survival curve anyway, this question might be considered first for removal or change. Questions asked in Gabor-Granger pricing method have some similarities to those required for modeling individual survival curves. Some further attempt to blend the PMS/NMS and Gabor-Granger pricing method could be also interesting.

Practical

On PSM, our bootstrapping exercises show PSM is highly unstable. We therefore encourage checking variability of results using bootstrapping. Some further investigation of the impact of response distributions on the intersecting points based on standard PSM could be beneficial. There could be a temptation to just ask and then report the average X2 and X3 without any consideration of revenue. Although not shown in this paper, that seems to overstate the optimal price.

On NMS, we question the purchase intention assumption for the "too cheap" price and thus call for caution and change as in its original form NMS likely results in highly overstated optimal price recommendations.

We hope with further validity checking that the two methods proposed are worth further consideration. We are also very eager to share the upcoming R package, hoping to benefit the market research community.

Ming Shan

References

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Carpenter, Bob, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32. https://doi.org/10.18637/jss.v076.i01.

Duursma, Remko A., and Brendan Choat. 2017. "Fitplc—an R Package to Fit Hydraulic Vulnerability Curves." *Journal of Plant Hydraulics* 4 (e002): 1–14.

Gelman, A., and J. Hill. 2000. "Data Analysis Using Regression and Multilevel/Hierarchical Models." 1st edition, *Cambridge University Press*.

Klein, Moeschberger, John P. 2003. "Survival Analysis—Techniques for Censored and Truncated Data." *Springer*.

Lipovetsky, Stan (2006). "Van Westendrop Price Sensitivity in Statistical Modeling." *International Journal of Operations and Quantitative Management*, 12(2), 141–156.

Miller, Newton, J. 1993. "A Market Acceptance Extension to Traditional Price Sensitivity Measurement." *Proceedings of the American Marketing Association Advanced Research Techniques Forum*.

Pinheiro, J. C., and D. M. Bates. 2000. "Mixed-Effects Models in S and S-Plus." *Springer*.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. 2021. "nlme: Linear and Nonlinear Mixed Effects Models." https://CRAN.R-project.org/package=nlme.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Van Westendorp, P, P. 1976. “NSS-Price Sensitivity Meter (Psm)—a New Approach to Study Consumer Perception of Price.” *Proceedings of the 29th ESOMAR Congress*, 139–67.

ESTIMATING WILLINGNESS TO PAY (WTP) GIVEN COMPETITION IN CONJOINT ANALYSIS

BRYAN ORME
SAWTOOTH SOFTWARE

This article describes how to simulate Willingness to Pay (WTP) in a more realistic and focused way than either the common algebraic approach or the two-product simulation approach. The common approaches don't consider competition and tend to overstate WTP. We recommend simulating product enhancements against a competitive set of alternatives (including the possibility of a None alternative) together with bootstrap sampling for estimation of confidence intervals around WTP. We introduce a generalizable and powerful extension called Sampling Of Scenarios (SOS) for estimating WTP that can be tailored to make certain detailed assumptions regarding the firm's product as well as competitive reactions in the marketplace. These new features are implemented in Sawtooth Software's desktop market simulator available within Lighthouse Studio and also as standalone Choice Simulator software.

INTRODUCTION AND MOTIVATION

Since the inception of conjoint analysis, researchers and their clients have sought intuitive ways to quantify the preference for attribute levels in monetary terms (e.g., Willingness to Pay or WTP). We should note that in economic studies, "paying" could involve other currencies such as time or travel distance and the approach we describe could be used to estimate WTP on other such currencies.

Historically, rather than reporting WTP, we at Sawtooth Software have preferred to quantify the impact of attribute levels on choice as changes in share of preference via sensitivity simulations. However, we also are frequently asked to deliver WTP. Over the last decade, these requests have only increased. Many consultants and other software packages compute WTP, but depending on the approach used, the monetary amounts are often overstated. The common approaches to WTP tend to overstate it, since they do not explicitly consider competition or the ability to opt out (choose the None). They also tend to average across respondents rather than focusing WTP more relevantly on respondents on the cusp of choosing the enhanced product features.

To promote better practice and more reasonable WTP results, we recommend procedures outlined below considering competition for estimating WTP via market simulation. These procedures are now available within our conjoint analysis market simulation software package, though researchers who have programming capabilities could implement them in other widely-used tools. Even though we've created easy-to-use software features to compute WTP, we caution researchers regarding pitfalls involved in interpreting WTP results, many of which we discuss below.

Failure to Account for Competition

In our opinion, failure to account for competitive alternatives is the main weakness in most commonly implemented WTP approaches and can lead to exaggerated WTP. In a 2001 paper later incorporated within the book, *Getting Started with Conjoint Analysis*, (Orme 2001, Orme 2004) we gave an example based upon the 1960s TV show, *Gilligan's Island,* illustrating how failure to account for competition can inflate WTP estimates. The cast is marooned on the island and a boat with capacity for two passengers appears on the scene ready to sell passage back to civilization to the highest bidders. The rich Mr. Howell appears willing to pay millions of dollars for passage for his wife "Lovey" and himself—until a second equally seaworthy boat appears offering the ride home for $5000. Mr. Howell of course chooses the $5000 option. The point of this illustration is that even though Mr. Howell is willing and able to pay over a million dollars, the availability and price of substitute goods in the marketplace means the firm (the boat) cannot capture this amount. If WTP is meant to represent the amount buyers are willing to pay the firm for enhanced features *in the current marketplace*, its calculation should account for competition including the None alternative (if available).

Two common approaches to WTP estimation do not consider competition or the ability to opt out and often can lead to inflated estimates of WTP:

1. The ***algebraic approach*** computes dollars per utile from the price function (the price utilities) and uses this to convert differences in utility between other attribute levels to monetary equivalents. This may be done at the individual level using HB utilities and the WTP estimates for the sample can be made more robust by taking medians (rather than means) across respondents.
2. The ***two-product market simulation approach*** simulates respondents choosing between just two versions of the product: one with and one without the enhanced feature. The price for the enhanced version of the product is adjusted upward via trial and error (or using our simulator's =SOLVEFORSHARE function) until the shares are distributed 50/50. The price difference that equalizes the shares of preference is taken as WTP. (Note: if the first-choice rule for simulating choice is used, the algebraic approach with medians and the two-product market simulation approach lead to identical WTP estimates. However, we generally recommend the logit share of preference approach.)

Simulation-Based WTP with Proper Competitive Context

Twenty years ago, Rich Johnson and this author recommended a *competitive set market simulation approach* for estimating WTP that accounts for competitive alternatives in the marketplace as well as the None alternative (Orme 2001). The approach involves first simulating[1] market choice for the unenhanced version of the

[1] For WTP estimation via simulations, we generally recommend using the Share of Preference (Logit equation) approach for estimating likelihood of choice for competing alternatives at the individual level, then averaging those results across respondents. First Choice, Randomized First Choice, and First Choice on the Draws could be used, but we would tend to prefer the Share of Preference approach operating on individual-level logit-scaled utilities for WTP estimation.

firm's product compared to a realistic set of competitive offerings and the None alternative. Next, we enhance the firm's product with a new feature and via trial and error (or using our simulator's SOLVEFORSHARE function). The increase in price that drives the share of preference for the firm back to the original base case share prior to feature enhancement is taken as the WTP. In real marketplaces, buyers are rarely limited to just one brand to obtain enhanced product features; they can select from many alternatives to achieve the same or compensating product benefits. Or, they can opt out. When competition is accounted for, WTP estimates are more realistic than methods that ignore competition.

HB-MNL and other MNL methods scale the utilities to be optimal for making predictions within the same context as the questionnaire's choice questions. Thus, if four alternatives were shown per question, the simulation predictions are more accurate when making predictions among a richer set of four alternatives rather than two alternatives as in the two-product market simulation approach.

We do not claim to be the first to propose using the market simulator with a competitive set of alternatives to find WTP as described in the previous paragraph. Rich Johnson, founder of Sawtooth Software, recommended it to me in the mid-1990s. Recently, I consulted via email with three very experienced conjoint researchers who were active in conjoint analysis in the 1980s (David Lyon, Keith Chrzan, and Joel Huber) and they agreed with Rich that estimating WTP using competitive simulation scenarios as I've described just seemed natural to do, given market simulation capabilities. They cannot recall being inspired regarding this by any specific article that detailed or originated the approach, as it just seemed common sense.

Upstream Steps to Improving WTP Analysis

Although not the subject of this article, I should note that hypothetical bias (e.g., respondents not spending real money in questionnaires or having to live with the consequences of their choices), interviewing the wrong people, and poor questionnaire design can inflate WTP estimates. We've outlined some ideas for improvement in these respects in our book, *Becoming an Expert in Conjoint Analysis* (Chrzan and Orme 2017). Noisy/bad data also can lead to exaggerated WTP and steps should be taken to remove respondents who appear to be answering randomly or completely ignoring price (Allenby et al. 2014). In extreme cases, we've found that data cleaning for speeders/random responders can remove as much as 50% of the data, though in our experience it's more typical to need to clean 15% to 25% of the sample.

Utility Constraints on Price

For most product categories that we'd expect to follow the law of supply and demand, we recommend constraining price to have negative slope (e.g., monotonically decreasing utilities as price increases) in estimation prior to computing WTP. Without price utility constraints, respondents with reversed price utilities could seem to choose a product with even higher likelihood as we increase the price. In some cases without price constraints, neither increasing nor decreasing the price of an enhanced product can drive its share back to the original share of preference prior to the product enhancement.

In such unusual cases, WTP via the competitive simulation approach doesn't yield a solution.

Proper Interpretation of WTP

As with other methods, our approach to WTP expresses monetary differences relative to a reference (base) level of an attribute. For example, if we've included three levels of speed (low, medium, and high) we consider a reference level (such as low speed) and estimate the value of the other two levels with respect to the reference level. For example, the relative WTP estimates may be:

Low speed	N/A (reference level)
Medium speed	$50 (relative to Low speed)
High speed	$120 (relative to Low speed)

Non-Additivity of WTP

Most approaches for estimating WTP for attribute levels focus on a single change in a product feature rather than a series of simultaneous feature improvements involving multiple attributes. A common error in interpreting WTP is to assume that WTP is additive across independent attributes. For example, if each of six features has an estimated WTP of $50 when *individually* and independently enhancing a base case product, it would be extrapolating beyond the assumptions of our WTP approach to conclude that the WTP for all six features *simultaneously* added to a base case product is 6 x $50 or $300.

Simply summing WTP values fails to account for diminishing marginal returns for cumulative product improvements (which is accommodated by the sigmoidal shape of the logit function). Summing WTP values also fails to consider increasing resistance due to buyers' budgetary constraints where increasing the price by cumulative amounts may very well push the utility function for price into a new region of the utility function that reflects greater price sensitivity. We can account for this by doing WTP analysis for multiple features taken simultaneously; it just requires simulating the firm's base case product followed by a version of the product enhanced by *multiple* features (again vis-à-vis relevant competition and the None alternative) and finding the indifference price via trial and error or using Sawtooth Software's automated SOLVEFORESHARE function.

New Advances in Simulating WTP for Conjoint Analysis

Next, we'll describe the main advances within Sawtooth Software's simulator for estimating WTP.

1. We've automated a search procedure for finding the change in price that leads to share of preference indifference (i.e., equality in preference shares) between enhanced and base case versions of the firm's product when those are placed in competition with other alternatives and the None.

2. Previously, there wasn't a straightforward way using our software to compute confidence intervals when using the recommended market simulation approach for computing WTP. We've implemented bootstrap sampling[2] to achieve confidence interval estimates of WTP.
3. For instances where the competitive set isn't well known or if the researcher wants to account for uncertainty in competitive reactions, we propose and introduce repeated Sampling Of Scenarios (SOS) with varying feature characteristics and pricing for both the firm's product as well as competitors.

Contrasts to Other WTP Approaches

The two common approaches to estimating WTP described earlier in this article (the *algebraic approach and* the *two-product simulation approach*) are unrealistic, failing to account for one or more of the following:

- The firm usually doesn't hold a monopoly on enhanced features; competitors can also provide the enhanced features, sometimes at a price lower than the firm hopes to charge.
- Competitors can provide other combinations of features or prices that may attract buyer preference despite the feature enhancements made by the firm.
- Competition doesn't necessarily remain static but may react to the firm's product enhancements in myriad ways, either rational or irrational.
- Consumers aren't forced to buy anything, but usually can opt out (choose the None alternative).
- WTP primarily should consider the respondents likely to switch to or away from the firm's product (those on the cusp of buying).
- WTP should depend on the firm's current positioning and price. A base case price of $300 could lead to a different WTP for an enhanced feature than a base case price of $400.

Some approaches to economic valuation of product features focus on estimating additional profits to the firm due to product enhancement, accounting for competitive reaction and assuming a game theory framework where the firm along with competitors are guided by profit maximization goals until achieving Nash Equilibrium. Such an approach usually requires knowing or assuming feature costs for each of the players (such that profit may be computed), which is nearly impossible to ascertain in most

[2] We've long advocated using HB estimation for conjoint models and for ease of implementation and efficiency for practitioners in the trenches we've relied upon summary point estimates rather than the granular draws to predict respondent choices. We admit this is a simplification that departs from a full Bayesian treatment for representing uncertainty and confidence bounds. Instead, we employ bootstrap sampling to estimate confidence intervals for the simulation-based WTP method. Another simplification is to conduct bootstrap sampling within the same HB estimation run. Formally, bootstrap sampling should be done with new HB estimation runs (a separate HB estimation for each bootstrap sample), rather than just bootstrap sampling the posteriors within the same HB estimation run. We compare results for those two approaches in Appendix B.

situations. Neither of the two competitive simulation approaches to WTP we describe below require knowledge of costs.

The Fixed Competitors Simulation Approach

When the competitive set is known or may be approximated, the researcher can specify the firm's base case product and price along with the fixed characteristics of competitors within a simulation scenario. An exhaustive set of competitors doesn't need to be specified; but the important players in the market should be represented. As is best practice for conjoint market simulations, we recommend using high-quality individual-level utilities from hierarchical Bayesian (HB) estimation, Mixed Logit, or similar estimation methods. To estimate WTP for features associated with the firm's base case product, we employ the preference share (indifference) approach which finds the change in price associated with a product enhancement that drives the share of preference for the firm back to its original preference prior to making the enhancement. When an attribute is not the focus of WTP estimation, we return it to its base case specification for the firm's product. We can repeat the simulation multiple times using bootstrap sampling to estimate confidence intervals for WTP of the features. We recommend at least 300 bootstrap samples for reasonable estimates of confidence intervals and 1000 or more bootstrap samples if high precision is desired.

The Sampling Of Scenarios (SOS) Approach

The Sampling Of Scenarios (SOS) approach is a useful generalized approach when there isn't certainty as to the base case product specifications or when there is uncertainty about competitor composition and reactions. What makes the SOS approach different from the fixed competitors approach is that we repeatedly sample among randomly selected competitive positioning as well as random variations in the firm's product for which we are estimating WTP. For each sampled scenario, we estimate WTP in the same way we've described above: through finding the equalization price for the enhanced product that sets its share back to the original share prior to enhancement.

Our approach to SOS[3] can involve more than just a generalized random selection of features. If specified, it can account for certain assumptions (constraints) regarding the firm's focus offering or the competitors' features and pricing. For example, we can assume the firm cannot change its brand or (optionally) the level of one or more other attributes. We can use a researcher-defined base price for the firm's offering, or we can generalize it to cover the entire price range. If the competitive offerings cannot include certain attribute levels, we can specify those exclusions (e.g., the competitors cannot assume the brand level associated with the firm). If the firm believes it can hold a

[3] Dave Lyon, the reviewer for this paper, suggested (and we agree) that the SOS approach could be used for general sensitivity analysis for attribute levels. This is where we specify competitive scenarios and observe how changes to the base case levels of the client's product affect the share of preference outcomes.

monopoly or patent on certain other feature(s), then the competitors can be prohibited from taking on such feature(s).

The reader may wonder how long it takes the software we've developed to perform WTP estimation for all attributes in a study for, say, 1000 Sampling Of Scenario draws. It takes around 5 to 45 minutes for the typical commercial CBC data sets we've experimented with. If performing bootstrap sampling with SOS draws, it can take 9 times as long as this if using the software's defaults, meaning a run that potentially takes a long lunch break to overnight to finish.

The SOS method seems robust to variation in the number of competitors used in the simulation scenario. (Appendix A)

Example Results for Conjoint Data Sets

We have tested WTP estimation using the new Sampling Of Scenarios (SOS) approach for nine conjoint data sets. All nine cases used HB estimation and we constrained price to be negative. For comparison, we also report the two common approaches for computing WTP that don't assume any competition (algebraic approach with medians[4] and two-product[5] simulation).

The results are fairly consistent across CBC datasets. The SOS approach tends to lead to the lowest estimates of WTP of the three methods we tried and also allows us to go beyond limitations of the other two approaches. For illustration, one of the datasets led to the following average WTP values:

	Algebraic Approach to WTP (medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
TV Data Set (n=382):	$106	$99	$79

WTP for this TV Data Set is fairly representative of the results across the nine datasets, where the SOS approach usually leads to the lowest estimates of WTP and the algebraic approach usually leads to the highest WTP estimates.

To add more color to the analysis and demonstrate the additional capabilities of the SOS approach to WTP, we can estimate WTP assuming an exclusive patent on a feature. Let's imagine our product has a patent on Channel Blockout technology. This means that we can specify in the software that the random draws of competitors cannot take on this level. In that case, WTP for Blockout technology increases from $56 in the general case where competitors can include this feature to $87 if we hold an exclusive patent.

[4] For the algebraic approach with medians, we generally used part-worth price coding and constrained price to be negative. We calculated the dollars per utile by referencing the lowest and highest price levels.

[5] For the 2-product simulation approach, we generally simulated the two products starting at a price point about in the middle or lower third of the price range. We felt this would better approximate the average price sensitivity across the price function than, for example, always starting the simulations referencing the lowest price.

WTP for Blockout with and without Patent

With Patent:	$87
Without Patent:	$56

As a second example of the strength of the SOS approach to WTP, we can isolate WTP for Blockout technology holding a level of a different attribute constant, such as brand. There are three brands in this old CBC study collected in the mid-1990s: JVC, RCA, and Sony. From past experience with this data set (via Latent Class analysis), we know that respondents who prefer Sony tend to be less price sensitive than those preferring the other brands. We can run WTP analysis using the SOS approach, where the WTP product always carries (in turn) the Sony, RCA, or JVC brands and the random draws of competitors take on the other two brands. The resulting WTP for Blockout by brand is:

WTP for Blockout by Brand

JVC	$54
RCA	$50
Sony	$66

Below, we report WTP results for all nine data sets, where we've indexed each WTP estimation to 1.0 for comparison.

TV Data Set (n=382):

	Algebraic Approach to WTP (medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Mono to Stereo	1.12	1.11	0.77
No Blockout to Blockout	1.09	0.99	0.92
No Picture-in-Picture to PIP	1.14	0.99	0.87
Column averages:	1.12	1.03	0.85

Cessna Airplane Data Set (n=539):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Brand1 to Brand4	1.18	1.10	0.71
A2L3 to A2L2	0.95	0.93	1.13
A3L1 to A3L2	1.05	1.03	0.92
A3L1 to A3L3	1.03	1.02	0.95
Column averages:	1.05	1.02	0.93

Phone Data Set (n=586):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Brand5 to Brand4	0.80	0.95	1.24
A3L1 to A3L2	1.07	1.07	0.86
A4L1 to A4L3	1.01	1.08	0.92
A5L4 to A5L1	0.87	1.04	1.09
A5L4 to A5L3	0.95	1.05	1.09
A6L4 to A6L1	1.03	1.05	0.92
Column averages:	0.95	1.04	1.01

INDIAA Data Set (n=1202):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Brand9 to Brand4	0.97	0.81	1.22
Brand9 to Brand5	1.23	0.78	0.99
Brand9 to Brand6	0.96	0.64	1.39
A2L6 to A2L4	1.23	1.12	0.65
A2L6 to A2L8	1.82	0.87	0.32
Column averages:	1.24	0.84	0.91

Cruise Line Data Set (n=600):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Carnival to Norw	0.83	1.29	0.88
Inside to Ocean	1.17	1.14	0.69
Older to Newer	1.02	1.08	0.90
11 days to 7 days	0.87	1.13	1.00
Column averages:	0.97	1.16	0.87

Cons Data Set (n=120):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
A1L1 to A1L2	0.75	1.10	1.15
A3L6 to A3L4	1.41	0.87	0.71
A3L6 to A3L5	1.20	1.05	0.74
Column averages:	1.12	1.01	0.87

Chspr Data Set (n=356):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Brand4 to Brand3	1.60	0.81	0.58
A2L2 to A2L3	0.96	0.99	1.05
A3L1 to A3L2	1.30	1.13	0.57
A4L2 to A4L1	1.79	0.79	0.42
A5L1 to A5L2	1.31	0.98	0.70
Column averages:	1.39	0.94	0.66

Flat Screen TV Data Set (n=951):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Vizio to Samsung	0.75	0.96	1.29
1080p to 4K	0.62	1.07	1.31
No HDR to HDR	0.91	0.98	1.11
60Hz to 120Hz	0.81	1.43	0.77
3HDMI to 4HDMI	0.86	1.25	0.90
Column averages:	0.79	1.14	1.08

Study1 Data Set (n=420):

	Algebraic Approach to WTP (Medians)	2-Product Simulation Approach to WTP	SOS Approach vs. 5 competitors
Brand3 to Brand1	1.32	1.05	0.63
Brand3 to Brand2	1.16	1.09	0.75
A2L3 to A2L1	1.04	1.05	0.92
A3L1 to A3L2	1.31	1.01	0.69
Column averages:	1.09	1.03	0.88

Across data sets, the Algebraic approach tends to lead to higher WTP estimates. Counting how many times each approach led to the *highest* estimated WTP, we find:

Algebraic Approach (Medians)	6 out of 9
2-Product Simulation Approach	3 out of 9
Sampling Of Scenarios (SOS) Approach	0 out of 9

Averaging across the indices for the nine studies, the summary relative WTP prices are:

Algebraic Approach (Medians)	1.09
2-Product Simulation Approach	1.03
Sampling Of Scenarios (SOS) Approach	0.88

Referring to the relative WTP price indices, we see that the SOS approach leads to WTP values that are on average 14% lower than the 2-product simulation approach and 20% lower than the Algebraic approach with medians.

Bootstrap Sampling for Confidence Intervals

We can develop confidence intervals via bootstrap sampling. Recall that we are using HB estimation, so we have individual-level utilities. Bootstrap sampling involves sampling with replacement (repeatedly) as many respondents as are in the original data set. Note that sampling with replacement means that some of the original respondents will not appear in a given bootstrap sample, and others will appear two or more times. As an example, the INDIAA data set has 1202 respondents. We can create hundreds of samples of that data set each involving 1202 respondents via bootstrap sampling. The standard deviation of the WTP values across those resamples provides an unbiased estimate of the standard error. Then, the mean +/- 1.96 times the standard error defines the 95% confidence interval range.

Here are confidence interval results using bootstrap sampling for the INDIAA data set:

INDIAA Data Set (n=1202):

	SOS Approach vs. 5 competitors	Standard Error	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Brand9 to Brand4	$45.0	$7.2	$30.8	$59.1
Brand9 to Brand5	$33.8	$3.4	$27.1	$40.4
Brand9 to Brand6	$10.6	$3.7	$3.4	$17.9
A2L6 to A2L4	$43.2	$4.2	$35.0	$51.5
A2L6 to A2L8	$4.5	$2.3	$0.1	$8.9

Conclusions

Willingness to Pay (WTP) has been challenging for the research community to get right. Although conjoint analysis gives us the right kind of data from choices within realistic-looking market scenario contexts, the common approaches to estimating WTP from conjoint analysis data have weaknesses. Those common approaches include the algebraic method and the 50/50 two-product simulation approach. Estimating WTP using conjoint market simulators that incorporate a realistic set of relevant and appropriate competition (including the None option) leads to more realistic and lower estimates of WTP. The market simulation approach to WTP considering competition focuses the estimation on respondents who are on the cusp of choice, rather than averaging results across respondents who may have relatively little interest in a given feature enhancement. The Sampling Of Scenarios (SOS) extension to the competitive simulation approach allows us to either generalize WTP considering all possible competitive reactions, or to incorporate specific assumptions involving exclusivity of brand name or feature enhancements (e.g., due to a patent). Our results across nine commercial CBC datasets show that the market simulations approach considering a rich set of competitors obtains WTP estimates on average 20% lower than the common algebraic approach.

AREAS FOR FUTURE RESEARCH

Individual-level utilities from HB estimation are derived from a mixture of individual-level and upper-level (aggregate) information. Thus, we recognize that bootstrap sampling among lower-level HB utilities from a single HB estimation run may understate the true sampling variability in the WTP estimates, because we haven't re-estimated the utilities using HB for each resample. Rather, we are just taking a resampling among the HB utilities that have been estimated just once leveraging the full respondent sample. Estimating HB for each bootstrap sample would add a very large amount of time (typically 3 to 10 minutes for each resample). For conjoint studies that have a healthy number of choice tasks relative to parameters to estimate, the individual-level utilities rely less on the upper-level information, so confidence intervals may not be understated by much. However, for sparse conjoint datasets (relatively few choice tasks per individual relative to parameters to estimate), the upper-level model can have a fairly large influence on the individual-level estimates. We compare bootstrapped WTP results for confidence intervals between our approach and a more complete treatment involving re-estimation of the HB utilities for each resampling in Appendix B. But, more work could be done across a variety of datasets.

Running HB with useful external covariates can enhance the heterogeneity across respondent utilities. Thus, using covariates may provide more accurate (and wider) WTP confidence interval estimates when using our bootstrap approach than using plain vanilla HB estimation without covariates. The more sparse the conjoint data, the more helpful covariates could be in reflecting appropriate heterogeneity. We hypothesize that with a healthy number of choice tasks per respondent (15 or more choice tasks) and a few useful covariates related to preference, confidence interval estimates may not change much whether estimating HB separately for each bootstrap sample or not.

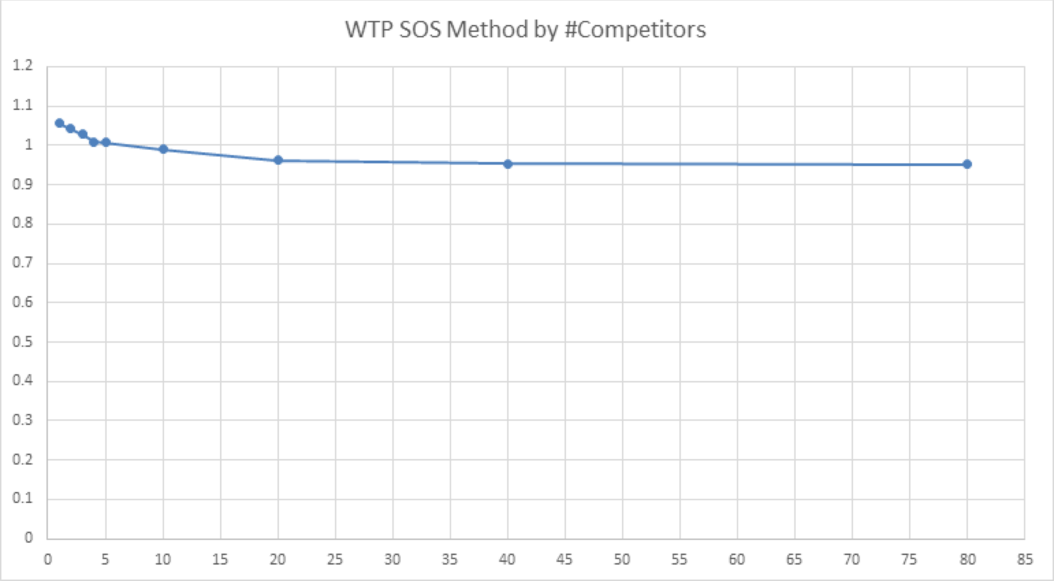
Bryan Orme

Appendix A

Robustness of WTP to Number of Competitors

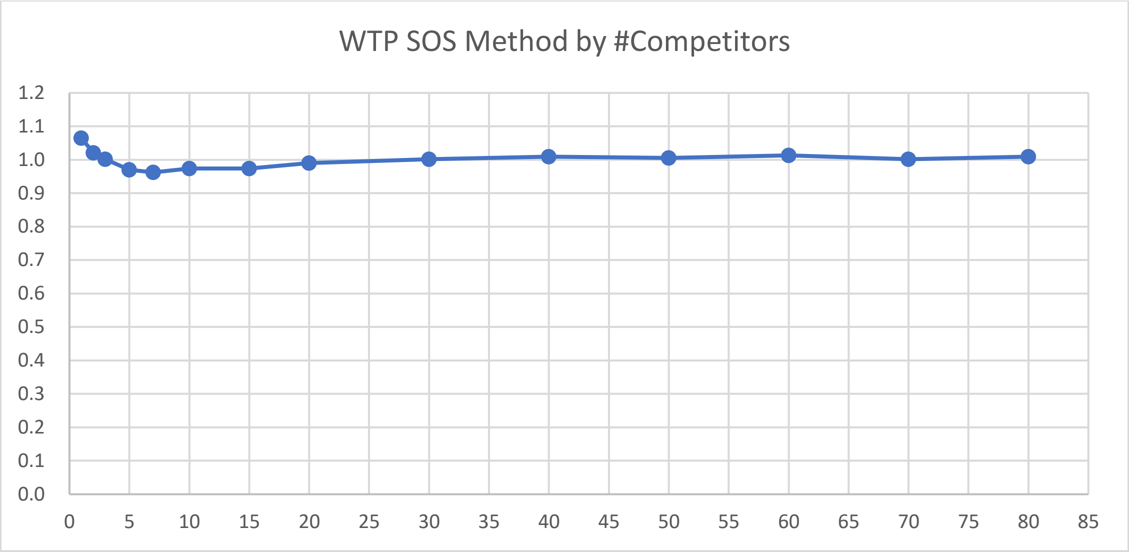
We have found that WTP estimates are fairly stable under different assumptions of number of competitors in Sampling Of Scenarios (SOS).

Using one of the CBC datasets cited earlier in this paper, we examined the robustness of WTP estimates for a given attribute level (relative to a reference level). Specifically, we varied the number of randomly drawn competitors in the simulation scenarios from 1 to 80. The results are shown below, where WTP is plotted on the Y Axis and indexed to 1.0 and number of competitors for SOS is represented on the X axis:



For this CBC dataset, the WTP with just one assumed competitor is about 7% higher than the WTP when 4 or 5 competitors are assumed. After about 20 assumed competitors, the WTP stabilizes with WTP about 5% lower than the WTP we found when using 5 assumed competitors.

We examined the same issue for a second CBC dataset and found that after five assumed competitors, the WTP results stabilized and did not change much at all:



We conclude that the SOS approach to estimating WTP is fairly robust to the number of assumed competitors. We expect results will vary somewhat depending on the data set and the number of levels in the attribute for which WTP is estimated.

For the software, we have set the default number of competitors for SOS to five. This would seem to strike a good balance between robustness of WTP results and speed.

APPENDIX B

Bootstrap Sampling within the Same HB Run vs. Separate HB Runs for Each Bootstrap Sample

To develop WTP confidence intervals, we've taken the simplification of bootstrap sampling the posterior utility estimates for respondents within the same HB estimation run (estimating HB utilities just once). Yet, since HB doesn't produce purely individual-level estimation (each individual's estimates are smoothed to some degree toward other members of the population), we may be understating the sampling distribution. Formally, it would be more appropriate to estimate WTP using independent HB estimation performed on each bootstrap sample. Unfortunately, this would dramatically increase the time requirement for obtaining confidence intervals via bootstrapping, making it prohibitive for practitioners.

How much our simplified bootstrapping approach leads to too-narrow estimates of confidence intervals depends on the degree of Bayesian shrinkage (to the upper-level model) in the posterior utility estimates. The more choice tasks per respondent, the less each respondent's utilities should be affected (via Bayesian shrinkage) by the surrounding population. Thus, for CBC datasets with relatively large numbers of choice tasks per respondent, we'd expect that the sampling distribution and resulting confidence intervals will tend to be larger and more accurate than for sparse data sets with relatively few tasks per respondent.

We compared WTP (using the simulation approach versus a set of fixed competitors) confidence interval estimates for two CBC datasets for two approaches: 1) simplified approach as used in Sawtooth Software's implementation where we use bootstrap sampling among the posterior individual-level utilities within the same HB run; 2) full approach where we conduct separate HB estimations within each bootstrap sample and use those utility estimates. For the comparisons, we used plain-vanilla HB estimation with no covariates (i.e., a single population assumption in the upper model). One CBC dataset was relatively sparse, with 8 choice sets and the other had a relatively large number of choice tasks (20).

We found with the 20-task CBC data set that the simplified approach produces confidence intervals about 25% narrower than the full approach that involves re-estimating HB utilities within each bootstrap sample. For the 8-task CBC data set, the confidence intervals were about half the width of the full approach.

We conclude that the simplified approach has the tendency to understate the width of the confidence interval for sparse CBC datasets using the plain-vanilla single population assumption in the upper model. If using our simplified bootstrapping approach and if more accurate confidence intervals for WTP estimates are needed, we recommend:

- Using CBC datasets where respondents answer at least 15 choice tasks,
- Using a few high-quality covariates (related to preference) in HB estimation to capture a more disperse representation of heterogeneity in the parameter estimates.

Taking these steps will allow use of the rapid simplified bootstrapping approach implemented in Sawtooth Software's market simulator while still achieving reasonably accurate estimates of WTP confidence intervals.

We should also note that using our software's Sampling Of Scenarios (SOS) approach tends to increase the standard error and its resulting confidence bound widths compared to using a set of fixed competitors. This isn't surprising, since sampling competitors adds another source of variability in the WTP estimates. In fact, with the SOS approach, the width of the confidence bounds may be overstated in many cases. Increasing the software's setting for number of Sampling Of Scenarios within each bootstrap loop (the default is 30) will reduce the degree of overstatement of confidence bounds. As software developers, we have to strike a balance between quality of results and time to compute. The default of 30 Sampling Of Scenarios iterations within each bootstrapping loop is one such judgement call.

REFERENCES

Allenby, Greg, Jeff Brazell, John Howell, and Peter Rossi (2014), "Economic Valuation of Product Features." *Quantitative Marketing and Economics*, 12:421–456.

Chrzan, Keith and Bryan Orme (2017), "Becoming an Expert in Conjoint Analysis," Sawtooth Software.

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

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

COMMENTS ON "ESTIMATING WILLINGNESS TO PAY GIVEN COMPETITION IN CONJOINT ANALYSIS"

DAVID W. LYON
AURORA MARKET MODELING, LLC

INNOVATIONS AND STRENGTHS OF ORME'S WTP APPROACH

The preceding paper by Bryan Orme presents a defensible and well-reasoned approach to answering the "willingness to pay" question from conjoint data. This is a problem familiar to many practitioners and the source of many problems (often in the form of laughably high estimates of willingness to pay) over the years. Bryan's solution is particularly satisfying because it is conceptually simple and doesn't involve much math. There is some computation involved, but in ways that many practitioners could program themselves if they had to.

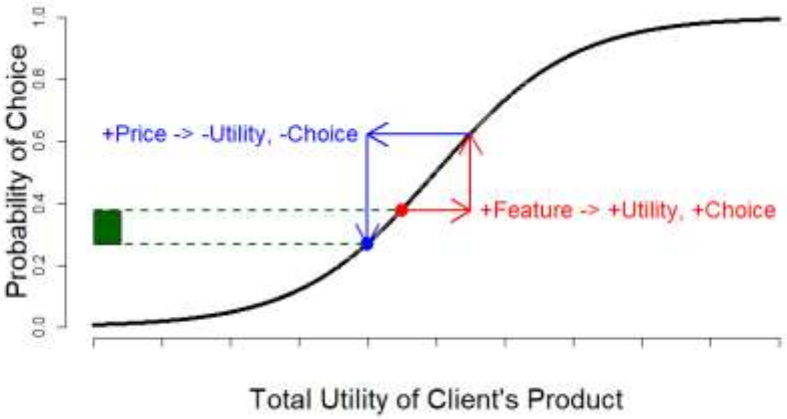
Bryan's approach recognizes or incorporates a number of important realities of WTP that some other approaches ignore:

- Respondents "on the cusp of choice" are who matter.
- WTP is not additive.
- WTP is not absolute, but dependent on the competitive context and the starting point for measurement.
- The managerial problem is to set a single price that works in the market, not to summarize the WTPs of heterogeneous respondents[1].

The Cusp of Choice

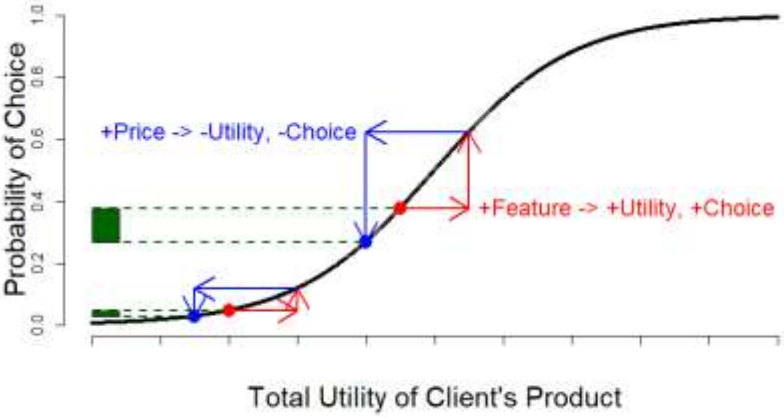
Consider the following graph of the logit curve that translates a respondent's utility for a product (x-axis) into her probability of choosing the product (y-axis), assuming some particular set of competitors.

[1] Allenby et al. (2014) address an even more on-point managerial problem: to maximize the client's profit in the context of full-knowledge competitive responses. It essentially involves setting a price that satisfies a Nash equilibrium in the "game" among competitors. This directly addresses the value of a feature, separately from WTP itself, and typically suggests a smaller price increase than any form of WTP analysis does. However, their approach places limitations on the formulation of the price utility and requires knowledge of all competitors' costs or margins.



If the respondent's utility is at the red circle on the curve and we then add a desirable feature to the client's product, utility increases and so does probability of choice, as shown by the red arrows. If we also increase the price, utility decreases, and so does probability of choice, in this case by more than the increases from the new feature. The green bar at the left highlights how much the probability of choice changed (about 10% or so). This respondent is "on the cusp of choice," or in the middle of the curve, where probability of choice is nearer 50% than to either extreme. As the steepest part of the curve, the middle is where smallish utility changes cause largish changes in choice probability.

Consider another respondent whose utilities for feature and price are identical to the first, but whose overall utility for the client product is far lower (perhaps due to a very low utility for the client's brand). In the graph below, this respondent is in the lower left corner, and even with identical price and feature utilities, his change in choice probability is far less. The respondent is very unlikely to choose the client's product in any event—they are not anywhere near "the cusp of choice."



By focusing on total simulated share, which is just the average of choice probabilities across all respondents, Bryan's approach implicitly places far more weight on the first respondent than on the second. It does this smoothly and elegantly, without arbitrary weighting or assignment of who matters and who doesn't, but simply as a natural by-product of the logit model analysis. This is a major strength of the approach.

Non-Additivity of WTP

Bryan's reminder that WTP is not additive is timely and useful. There are many stories from the 1980s of analysts telling clients to add every possible feature and triple their price and expect market success. Unfortunately, that continued well past the 1980s. This is a common and concrete special case of the context issue noted next.

Context-Dependence and Sampling Of Scenarios

WTP is not an absolute value, but depends on the client configuration taken as a starting point, and on the competitive context. In terms of the logit curves illustrated above, respondents' starting points depend on how the client is configured to start with. Even more obviously, they depend on the competition. WTP for a feature exclusive to the client will clearly be higher than if the same feature is offered by one or more competitors.

Occasionally, the right client starting point and competitive context is obvious. For example, pharmaceutical clients may have a definite target profile for the product and the competition is often well-known, with attributes that are invariant because of either chemistry or regulation. But more often, there is no natural starting or reference point for WTP or any other kind of sensitivity analysis.

The Sampling Of Scenarios (SOS) approach is an excellent way to address this situation. It removes any need to make an arbitrary choice of starting point. The

Sawtooth Software implementation that allows for prohibition of some levels for some competitors makes it particularly useful in practice by providing a way to specify the fixed parts of the situation, when we know them, while still randomly sampling the uncertain parts of the competitive situation.

The basic SOS idea should prove useful in all sorts of sensitivity analyses, not just in WTP.

The Right Managerial Problem

What underlies both the cusp of choice concept and the context-dependence concept is the adoption of the viewpoint of a price-setting manager. The fundamental WTP problem is to find a price for a feature that will maintain overall market share. (This could of course be generalized to such ideas as finding a price that would produce an x% gain in market share.)

Simulating the effects of adding a feature and increasing the price is the direct way to address that managerial problem. Too many other approaches are essentially statistical summaries, trying to find some useful way to make use of individual-level WTP calculations. But why do we care about things like algebraic equalization for those who won’t buy anyway, or who will (almost) always buy? Why do we care *who* buys, except to the extent it leads to better aggregate simulation results? For WTP purposes, we shouldn’t!

Starting with the business problem and working from there is usually a good idea, and Bryan’s approach does exactly that.

Bootstrapping for Confidence Intervals? Why Not the HB Posterior?

It is good to see confidence intervals (CIs) being explicitly considered. Too often, they are ignored until a client asks (and too few do) and then dismissed with worries about the difficulty of calculating them. It is commendable that Bryan and Sawtooth Software incorporated CI calculations from the beginning.

But, is bootstrapping the right way to get CIs? It is a solid technique, but as used here it accounts *only* for the respondent sampling variability in the results. The uncertainty in each respondent’s utilities (some might say *within* each respondent’s utilities) is ignored when bootstrapping on posterior means. The fundamental result of any Bayesian analysis is the posterior distribution. In our case, the HB posteriors tell us how well-determined each respondent’s utility estimates are. This information is being ignored when we use just the posterior means. Bootstrapping on the posterior means does nothing to restore that lost information[2].

Instead, we could run the share-equalizing price search at the heart of Bryan's method for each of, let's say, 1000 random draws from the HB posteriors and use their variance to generate the confidence intervals. Doing so would involve no more computation than working with 1000 bootstrap samples, so the workload and timings would be the same.

Working from the HB draws, however, would incorporate *all* the sources of uncertainty in the model, not just the respondent-sampling variance. Using the draws would be much more in the spirit of Bayesian analysis, where "the posterior answers all questions," as opposed to using bootstrapping as an add-on after simplifying the hierarchical Bayes analysis down to posterior means.

In my own experiments with a single study (with 12 tasks per respondent, 10 model parameters, no covariates, and 711 respondents), the CIs for 6 different feature changes estimated from HB draws[3] were as much as 2.6 times as wide as those estimated from bootstrapping, depending on the feature, the exact method of CI construction, and whether SOS was used, with most being at least 50% wider. These are substantial differences, implying that the bootstrapping process tends to yield confidence intervals that are far too optimistic.

In short, I believe the use of HB draws to replace the bootstrapping should be seriously considered. The results would be more rigorous, and the computational effort not much different[4].

Confidence Intervals with Sampling Of Scenarios

The SOS process adds additional variance (thus, widens the confidence interval) because of the variation in results for different scenarios. The current Sawtooth Software implementation averages results for 30 (or so) scenarios for each of 1000 (or so) bootstrap replications, and then calculates CIs based on the variance of those 1000 averages.

This unfortunately ignores most of the variance added by SOS. A better approach would be to calculate the variance of the 30 SOS results for each bootstrap replicate,

[2] As Bryan also notes, the implementation skips the HB re-estimation for each replicate that an ideal bootstrap implementation would involve. There are good practical reasons for that, but he sees as much as 25% to 50% of the total variance from the ideal being "lost" because of that simplification.

[3] This commenter used draws from the lower-level model. The upper-level draws could be used in very similar fashion.

[4] With current Sawtooth Software programs, there may be somewhat more human effort in using HB draws, in that the draws must be captured and managed, which is not common in everyday practice.

average those variances over the 1000 bootstraps and then add the net result to the overall variance estimated from the bootstrap means. The variance of the means is the inherent respondent/model variance (not controllable once data is collected); the average variance within a replicate is that due to SOS, which can be decreased, if felt to be too large, by re-running with more than the default 30 scenario samples.

In this commenter's experiments (on the same study mentioned in the previous subsection), accounting for the variance due to SOS with 30 scenarios increased standard errors and confidence interval widths by 2% to 22%, with many values around 12%. This seems reasonably small and very practical. (Those who find it too large could cut the increase in half by using four times as many SOS samples, which would correspondingly take four times as long to run.) But it really should be accounted for.

If CIs are calculated from HB draws as suggested in the foregoing subsection, exactly parallel issues apply for SOS: the SOS variance should be accounted for, but that can be done in computationally equivalent ways.

The variance added by SOS is a smaller issue than the question of bootstrapping vs. HB draws, but it would be easy (and appropriate) to account for in any future software update.

David W. Lyon

Reference

Allenby, Greg, Jeff Brazell, John Howell, and Peter Rossi (2014), "Economic Valuation of Product Features." *Quantitative Marketing and Economics*, 12:421–456.

FILTER CBC: A NEW APPROACH TO MIMIC THE ONLINE SHOPPING EXPERIENCE

MARCO HOOGERBRUGGE
MENNO DE JONG
KEVIN LATTERY
KEES VAN DER WAGT
SKIM

1. BACKGROUND AND INTRODUCTION

Many markets are hugely fragmented when it boils down to the SKU level. With, say, 10 brands the market seems to be quite simple, but if every brand offers, say, 10 different sizes or variants it becomes a different story. A respondent may prefer a certain SKU, but in certain scenarios they may switch to either a different SKU of the same brand under certain circumstances, or switch to a more similar competitor SKU. It becomes very difficult to assess which direction they are really taking when every SKU only has a tiny share of the market, whereas those sorts of details are awfully important to clients. That is most clearly visible when you look at the distribution of an individual respondent's shares of preference among the SKUs: it is usually greatly scattered among a substantial number of SKUs (even while a clear majority of SKUs show 0% share of preference). This paper is about a certain approach hoping to improve the predictions in such fragmented markets and is about measuring the improvement by comparing the predictions with real buying behavior.

The paper builds upon an earlier presentation by Menno de Jong and Lois van der Molen, delivered at the European Sawtooth Software Conference of September 2020. It is about an experiment we did in Spring-Summer 2020 which consisted of two very different components:

1. "External validity." We did not only do a choice exercise to predict what respondents would do, but we also checked a few months later what the respondents *had* bought. So, it was a two-wave survey with the same respondents. This way we could compare our predictions with the real choice rather than with holdout tasks which are (also) merely survey instruments.
2. "Filter CBC." In the first wave, we did the choice exercise in two different styles. One was the traditional CBC style; the other, called Filter CBC, was much more like a website where people could filter and sort based on product attributes before they made their choice.

The two components were not accidentally combined in one experiment. We suspected that Filter CBC would deliver better predictions because it allows respondents to make a choice between a much larger (i.e., realistically larger) set of products than in a traditional CBC exercise. But there is no way to establish the better performance by means of holdout tasks, because the style of the holdout tasks would heavily influence the outcome of the comparison. If the holdout tasks were more like traditional CBC tasks, they would probably assign traditional CBC tasks as the winner. If the holdout tasks were more like Filter CBC tasks, they would probably assign Filter CBC as the winner. The only objective way to compare is to have "outside the experiment" data, i.e., real buying behavior.

This paper consists of three parts. Part 2 provides more detail about the external validity component. Part 3 provides more detail about the Filter CBC component. And finally, part 4 goes in depth about different utility estimation methods for Filter CBC.

2. External Validity Assessment

In May 2020 we fielded the first wave of the experiment, with the choice exercises, and built a simulator with the current market so that for each respondent we could make a prediction. The product category was mobile phone subscriptions, and we had 6 attributes: brand, data, minutes, contract period, 4G/5G and price. Note that 4G/5G was added to the survey but did not play a role in the predictions because all products still had 4G at the time. The simulator that we constructed for our predictions consisted of 139 products. The country of fieldwork was the Netherlands. The number of respondents in the first wave was 1426, recruited based on the criterion that they would soon take a new subscription.

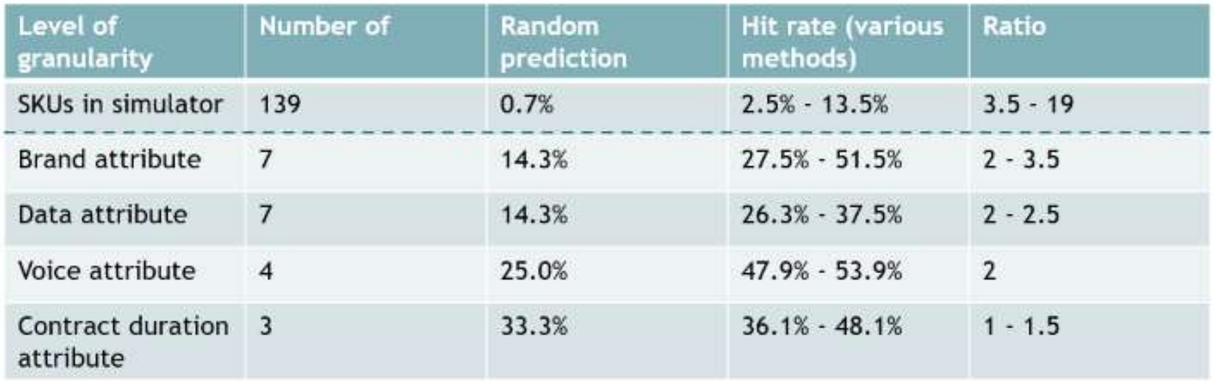
In July we managed to recontact about 2/3 of these respondents, of which 499 respondents qualified for the comparison because they had taken a new mobile phone subscription. Or to be more precise: (only) 21% had chosen a new contract with a new provider, (as much as) 42% had chosen an alternative contract with the same provider and (as much as) 37% had simply renewed the same contract. This alone is enormously important for our practical work because "staying with current brand" and "staying with current product" are phenomena that we often see in conjoint predictions to some extent, but seldom as extreme as these actual buying figures are telling here. In theory, a high utility for every respondent's current brand should increase the chance that we predict the same brand for many respondents. But in practice, in a survey setting, respondents switch brand easier than they do in real life.

In this second wave we asked respondents which new mobile plan they had taken with which provider, and in addition we asked a few extra questions. We also told respondents what we predicted for them, and if that was different, we asked an open question why respondents thought it was different. In advance we had expected that respondents would answer things about "external" factors here, i.e., reasons that cannot be captured by a conjoint model. To some extent that happened, because respondents mentioned they got an extra discount when prolonging their current subscription, which we cannot capture in the conjoint prediction. But we were mostly wrong in our expectations, because respondents mostly answered things like "I don't need that much data," "this brand is too expensive," i.e., they were mostly referring to factors that were *supposed* to be captured by our conjoint model. This implies that—somehow — there must be ample room for improvement for conjoint predictions.

As said, the second wave was aimed at collecting an objective measure of predictive validity. Now, the measure we got was *objective* for sure, but it was not always an *ideal* measure. Even data about real buying behavior sometimes have issues, apparently. For starters, a few dozen respondents had chosen another brand than we had in our conjoint grid, so we decided to omit these respondents entirely from the analysis. In addition, a few dozen respondents had chosen a brand-data-minutes combination that did not exist according to the website of the brand (i.e., it was a combination that we could not include in the simulator either). We decided to omit these respondents from the analysis too. Finally, the reported price of the brand-data-minutes price combination was often different than listed on the website of the brand. It was mostly higher and that might well have been because respondents added the monthly cost of the handset to the price (even while we specifically asked them not to do that). If the mentioned price was lower on the other hand, it could have been the case that they got a special discount for renewing their existing subscription. We decided to *keep* these respondents in the analysis and largely focus our comparisons on the four non-price attributes (brand, data, minutes, and contract period).

The goal was to compare the wave 1 predictions and the wave 2 actual data. The match turned out very favorable (see Exhibit 1). We had not expected otherwise, by the way 😊. On an attribute level the hit rate was mostly around 2 times the theoretical random hit rate, but at individual product level, the hit rate was 3 to 19 times the theoretical random hit rate, and that is what matters most. This is a broad range, because we tested many estimation methods (as will be detailed in part 5).

Exhibit 1: Comparison of Hit Rate versus Random Prediction



Level of granularity	Number of	Random prediction	Hit rate (various methods)	Ratio
SKUs in simulator	139	0.7%	2.5% - 13.5%	3.5 - 19
Brand attribute	7	14.3%	27.5% - 51.5%	2 - 3.5
Data attribute	7	14.3%	26.3% - 37.5%	2 - 2.5
Voice attribute	4	25.0%	47.9% - 53.9%	2
Contract duration attribute	3	33.3%	36.1% - 48.1%	1 - 1.5

At the same time, a half-full glass is also a half-empty glass, so it is also important to mention that there is still a huge amount of prediction performance to gain, since the hit rates are not getting anywhere near 100% or even near 70%. This relates also to the earlier comment we made about the open answers in wave 2. For example, we predicted the right brand for only half of the respondents (in the best case), and the right amount of data for only a third of them (in the best case).

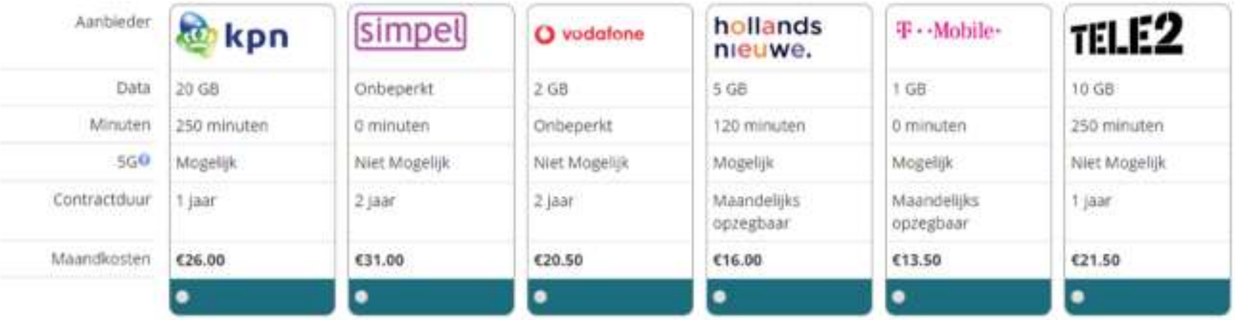
On top of that, for brand (where we had 3 "premium" brands and 4 "cheap" brands) we see a peculiar phenomenon: *if* the conjoint model predicted a "premium" brand for a respondent, then they surely chose a premium brand in real-life—and then it is very likely the exact brand that was predicted! In other words, here the utilities reflect real life behavior well. But *if* the conjoint model predicted a "cheap" brand for a respondent, then they might well have taken a premium brand after all, and often they chose the brand that they currently had. The "cheapest" brand in particular was predicted quite a lot but chosen very little in reality. So, in the latter case the brand utilities do not reflect real life behavior well enough.

One practical idea to improve predictions further, outside of the conjoint modeling, is to favor respondents' current brand more in the predictions. That was presented in the earlier European Sawtooth Software Conference. The extent to which this needs to happen may vary from one study to the next, however. Another idea is to weight respondents by their predicted brand (so not their current brand): respondents with extremely price sensitive survey behavior will get a very low weight.

3. Filter CBC and the Comparison with Traditional CBC

The traditional CBC that we used in this study wasn't fully traditional because it had one important special feature: the price was determined by summed pricing, i.e., as a sum of attribute level-specific price amounts with some random variation (like in ACBC). Especially the amount of data is driving the price of mobile subscriptions a lot, so this was necessary to keep prices realistic. An example of how CBC looked is shown in Exhibit 2.

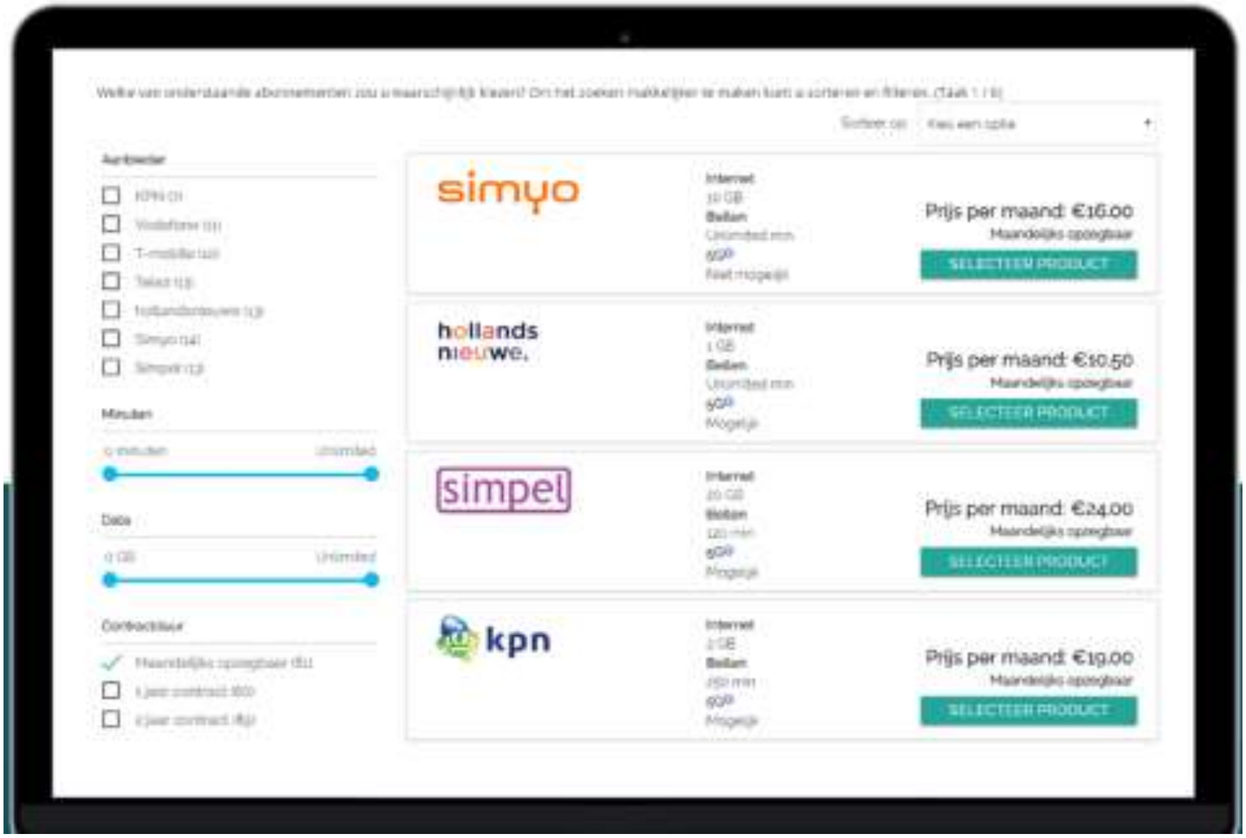
Exhibit 2: Example of Traditional CBC Task



Aanbieder	kpn	simpel	vodafone	hollands nieuwe.	T··Mobile·	TELE2
Data	20 GB	Onbeperkt	2 GB	5 GB	1 GB	10 GB
Minuten	250 minuten	0 minuten	Onbeperkt	120 minuten	0 minuten	250 minuten
5G	Mogelijk	Niet Mogelijk	Niet Mogelijk	Mogelijk	Mogelijk	Niet Mogelijk
Contractduur	1 jaar	2 jaar	2 jaar	Maandelijks opzegbaar	Maandelijks opzegbaar	1 jaar
Maandkosten	€26.00	€31.00	€20.50	€16.00	€13.50	€21.50
	○	○	○	○	○	○

Filter CBC looks a lot fancier, and more in line with how we nowadays buy products on websites. An example of how Filter CBC looks is shown in Exhibit 3. It consists of attribute levels to be filtered in the left panel, and product concepts displayed vertically in the middle and right panel. Additionally, in the right top corner there is the possibility to sort the concepts with respect to a certain attribute ("sorteren op").

Exhibit 3: Example of Filter CBC Task



Welke van onderstaande abonnementen zou u waarschijnlijk kiezen? Om het zoeken makkelijker te maken kunt u sorteren en filteren. (Taak 1 / 8)

Sorteer op: Kies een optie

Aanbieder

- □ KPN (1)
- □ Vodafone (11)
- □ T-mobile (12)
- □ Tele2 (13)
- □ Hollandsnieuwe (13)
- □ Simyo (14)
- □ Simpel (13)

Minuten

0 minuten — Unlimited

Data

0 GB — Unlimited

Contractduur

- ✓ Maandelijks opzegbaar (81)
- □ 1 jaar contract (80)
- □ 2 jaar contract (83)

simyo
Internet
10 GB
Bellen
Unlimited min
5G
Niet mogelijk
Prijs per maand: €16.00
Maandelijks opzegbaar
SELECTEER PRODUCT

hollands nieuwe.
Internet
1 GB
Bellen
Unlimited min
5G
Mogelijk
Prijs per maand: €10.50
Maandelijks opzegbaar
SELECTEER PRODUCT

simpel
Internet
20 GB
Bellen
120 min
5G
Mogelijk
Prijs per maand: €24.00
Maandelijks opzegbaar
SELECTEER PRODUCT

kpn
Internet
2 GB
Bellen
250 min
5G
Mogelijk
Prijs per maand: €19.00
Maandelijks opzegbaar
SELECTEER PRODUCT

Underlying, there is a two-attribute design of 100 products per choice task (a random selection of the 139 that we have in the simulator) and price. The product attribute has been expanded to data, minutes, contract duration and speed in line with the real-life specifications. Lastly price, like in CBC, is calculated as summed pricing including some random variation.

The products shown in the choice task are taken from the products in the design insofar they fit the respondent's filter criteria. In this case only monthly contracts are filtered in by the respondent, so all concepts shown have monthly contracts. As soon as the respondent changes filter criteria, the product concepts are being updated as well. The concepts fitting the filter criteria are either shown in random order, or in order of a certain attribute if the respondent has used that option. Respondents can scroll down to look at more product concepts than just the four that are shown on top.

The obvious advantage of the style of the Filter CBC tasks is that it is more "realistic" to respondents (at least for certain product categories). There is also a different, methodological, advantage in play. What we do here is *expand the choice set* from 6 to 100 concepts. In the Filter CBC data processing, we assume that 99 concepts have been rejected even though respondents will not have explicitly evaluated each of those 99 concepts individually. By making this assumption, we supposedly will have much more detailed insight in what respondents want to have in the real market.

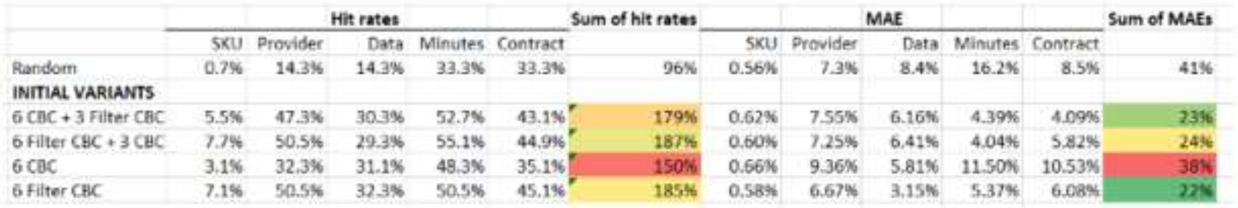
The way we implemented traditional CBC and Filter CBC in the survey is an important thing to discuss as well. An AB test would not as easily provide significant differences because of the sampling error involved. So instead, we chose a within-respondent design and requested all respondents to take both traditional CBC and Filter CBC tasks. But on the other hand, we would not want to burden the respondents too much, so we gave each of them 6 traditional CBC tasks and 6 Filter CBC tasks, in random order. Interestingly, respondents did not have a significant preference for either of the two styles in terms of user experience. The average score was 7 on a 0–10 scale. Some respondents mentioned advantages of Filter CBC ("convenient filters" and "almost like real"), others mentioned disadvantages ("not enough overview," "too many choices"). For traditional CBC, people mentioned that there often were no suitable choices available for them, so that is an implicit compliment for Filter CBC, while on the other hand other people like traditional CBC better as it is quicker to answer.

Initially, we presumed that running HB independently on just traditional CBC 6 tasks and on 6 Filter CBC tasks would not be a good idea because then we would have too few data points per respondent. So, we decided to go forward with some mix and match and we compared:

a. 6 traditional CBC tasks + 3 Filter CBC tasks
b. 3 traditional CBC tasks + 6 Filter CBC tasks

As a matter of deduction, we would argue if (b) performs better than (a) then Filter CBC performs better than traditional CBC. This is the way we presented results at the European Sawtooth Software Conference, however with the rather disappointing result that (a) and (b) did not perform a lot differently from each other. Only later we found that we were mistaken in our assumptions. In Exhibit 4 we compare not only a) and b) but also just 6 Filter CBC and just 6 traditional CBC tasks. Here it appears that (a) and (b) and just 6 Filter CBC tasks perform much better than just 6 traditional CBC tasks. In other words, 3 extra Filter CBC tasks add a lot to 6 traditional CBC tasks, while 3 extra traditional CBC tasks do not add anything to just 6 Filter CBC tasks. So, this is clearly an indication that Filter CBC outperforms traditional CBC.

Exhibit 4: Comparison of CBC and Filter CBC in Terms of Hit Rate and Mean Absolute Error, Taking the Real-Life Purchase of Respondents as the Benchmark



	Hit rates					Sum of hit rates	MAE					Sum of MAEs
	SKU	Provider	Data	Minutes	Contract		SKU	Provider	Data	Minutes	Contract	
Random	0.7%	14.3%	14.3%	33.3%	33.3%	96%	0.56%	7.3%	8.4%	16.2%	8.5%	41%
INITIAL VARIANTS												
6 CBC + 3 Filter CBC	5.5%	47.3%	30.3%	52.7%	43.1%	179%	0.62%	7.55%	6.16%	4.39%	4.09%	23%
6 Filter CBC + 3 CBC	7.7%	50.5%	29.3%	55.1%	44.9%	187%	0.60%	7.25%	6.41%	4.04%	5.82%	24%
6 CBC	3.1%	32.3%	31.1%	48.3%	35.1%	150%	0.66%	9.36%	5.81%	11.50%	10.53%	38%
6 Filter CBC	7.1%	50.5%	32.3%	50.5%	45.1%	185%	0.58%	6.67%	3.15%	5.37%	6.08%	22%

4. Alternate Estimation Methods for Filter CBC (Introduction)

In the previous sections we only considered two sources of data: the 6 CBC tasks and the 6 Filter CBC tasks. However, we also have data from the *filtering process* in Filter CBC, where respondents tell us which attribute levels they reject and thereby create a consideration set. We can also use this third source of data for analysis. But it is less straightforward if and how to do that exactly. This section is about different ways of including that filter process information in the modeling. Note that the methods we discuss are not available in Sawtooth Software. We used custom programming in R and Stan.

4.1 Screen Alternatives

One way to think about the filtering process is that the respondent is screening alternatives. For each respondent, each alternative has a utility. The respondent compares this utility to a certain threshold. Alternatives that remain after filtering have a utility greater than a certain threshold, while those that were filtered out have a lower utility than the threshold.

This approach is consistent with the method developed by Lattery (2010) for anchored MaxDiff. There are two approaches in which this can be specified. Lattery recommends creating two additional tasks, one where the items beat the threshold and another where the items lose to the threshold.

For the "bad" items (filtered out) we code them in the obvious way, listing each of the bad items and a threshold, where the threshold wins:

Filtered Out	**Response**
Bad$_1$	0
Bad$_2$	0
Bad$_n$	0
Threshold	1

For the good items (filtered in) the coding is more complex. Here threshold < items, which is equivalent to -1*threshold > -1 * items. So the coding looks like this:

Filtered In	**Response**
Good$_1$ *-1	0
Good$_2$ * -1	0
Good$_m$ *-1	0
Threshold *-1	1

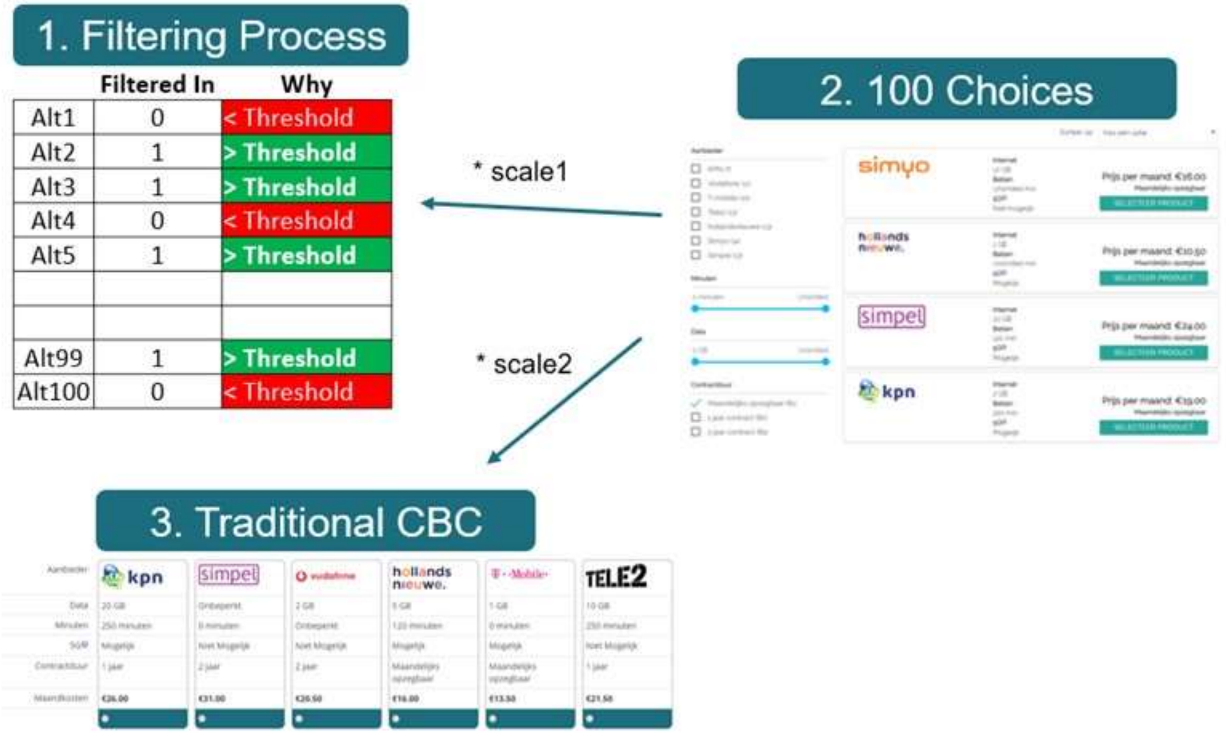
In general, there will be a different number of alternatives in the good and bad set. It is even possible that the bad set is empty, if the respondent applied no filter at all.

While Lattery prefers this two-task augment, others find it more intuitive to apply a set of binary tasks. In our study with 100 alternatives this requires adding 100 additional tasks. Each task has two alternatives. Threshold beats a bad item, and loses to a good item:

Filtered Out	**Response**		**Filtered In**	**Response**
Item	0		Item	1
Threshold	1		Threshold	0

Kees van der Wagt did the analysis for this method and applied the coding as binary tasks.

The set of binary tasks poses two problems. First, we have a very different scale in the binary tasks versus the conjoint where we show 100 alternatives. Indeed, we actually have two scale factors, one for the binary tasks and another for the traditional CBC:



1. Filtering Process

	Filtered In	Why
Alt1	0	🟥 < Threshold
Alt2	1	🟩 > Threshold
Alt3	1	🟩 > Threshold
Alt4	0	🟥 < Threshold
Alt5	1	🟩 > Threshold
Alt99	1	🟩 > Threshold
Alt100	0	🟥 < Threshold

2. 100 Choices

* scale1

* scale2

3. Traditional CBC

Aanbieder	kpn	simpel	vodafone	hollands nieuwe.	T-Mobile	TELE2
Data	20 GB	Onbeperkt	2 GB	5 GB	1 GB	10 GB
Minuten	250 minuten	0 minuten	Onbeperkt	120 minuten	0 minuten	250 minuten
5G/4G	Mogelijk	Niet Mogelijk	Niet Mogelijk	Mogelijk	Mogelijk	Niet Mogelijk
Contractduur	1 jaar	2 jaar	2 jaar	Maandelijks opzegbaar	Maandelijks opzegbaar	1 jaar
Maandkosten	€26.00	€31.00	€26.50	€16.00	€13.50	€21.50

If one cannot apply separate scale factors then we highly recommend the two task supplement. In that case there will typically be many more alternatives in each task, so there will not be such a strong difference in scale.

In addition to the difference in scale, we are also adding 100 tasks for each of the 6 filtering processes. It is worth noting that respondents could change the filter each task, and many did. So, we added 100 * 6 = 600 tasks. These 600 tasks for each respondent would clearly dominate the analysis. So we had to weight each of the 600 supplemental tasks down. Kees van der Wagt experimented with many different sets of weights. His results are shown later.

If one cannot apply separate scale factors or weights for tasks, then we highly recommend the two-task augment. In that case there will typically be many more alternatives in each task (all the good or bad items), so there will not be such a strong difference in scale. And adding two tasks is far less dominating, so weighting them down is probably not needed.

4.2 Screen Levels of Attributes

The second method (preferred by Lattery) does not treat the filtering process as screening alternatives. Rather the respondent is viewed as screening levels of attributes. The respondent is telling us that these levels of these attributes are unacceptable.

Here our analysis is analogous to the Gillbride and Allenby (2014) paper on conjunctive screening rules. Using the attributes from our case study, conjunctive screening defines those items remaining (after filtering) for a respondent as those for which:

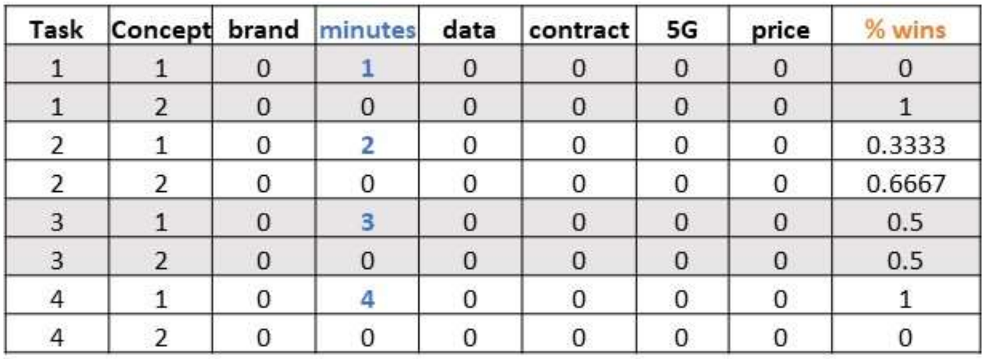
Utility of brand > brand threshold &
Utility of minutes > minutes threshold &
Utility of data > data threshold &
Utility of contract > contract threshold

Mathematically, Gillbride and Allenby use a binary indicator function for each condition and multiply these to determine whether the composite alternative is filtered in or not. We used probabilities rather than binary indicators:

Prob(Utility of brand > brand threshold) *
Prob(Utility of minutes > minutes threshold) *
Prob(Utility of data > data threshold) *
Prob(Utility of contract > contract threshold)

And each of the probabilities is defined by a binary logit: $e^{U}/(e^{U} + e^{\text{Threshold}})$.

For example, we had 4 levels of a minutes attribute. We would add 4 partial profile tasks, comparing a specific level of the minutes attribute with a 0 vector that represented the threshold for minutes. The other attributes are 0. So the 4 tasks would look like this:



Task	Concept	brand	minutes	data	contract	5G	price	% wins
1	1	0	1	0	0	0	0	0
1	2	0	0	0	0	0	0	1
2	1	0	2	0	0	0	0	0.3333
2	2	0	0	0	0	0	0	0.6667
3	1	0	3	0	0	0	0	0.5
3	2	0	0	0	0	0	0	0.5
4	1	0	4	0	0	0	0	1
4	2	0	0	0	0	0	0	0

At the far right we have the percentage of time each level was filtered in across the 6 tasks. Respondents changed their filters across the 6 tasks. The first task shows that level 1 of minutes was always filtered out—it was always unacceptable. Level 2 was filtered in 1/3 of the time, and level 3 50% of the time. The levels of minutes were ordinal, so level 4 (unlimited minutes) was always filtered in. Note that the table above shows the levels for minutes. We would still code the minutes as a categorical variable with 4 levels since it is not linear.

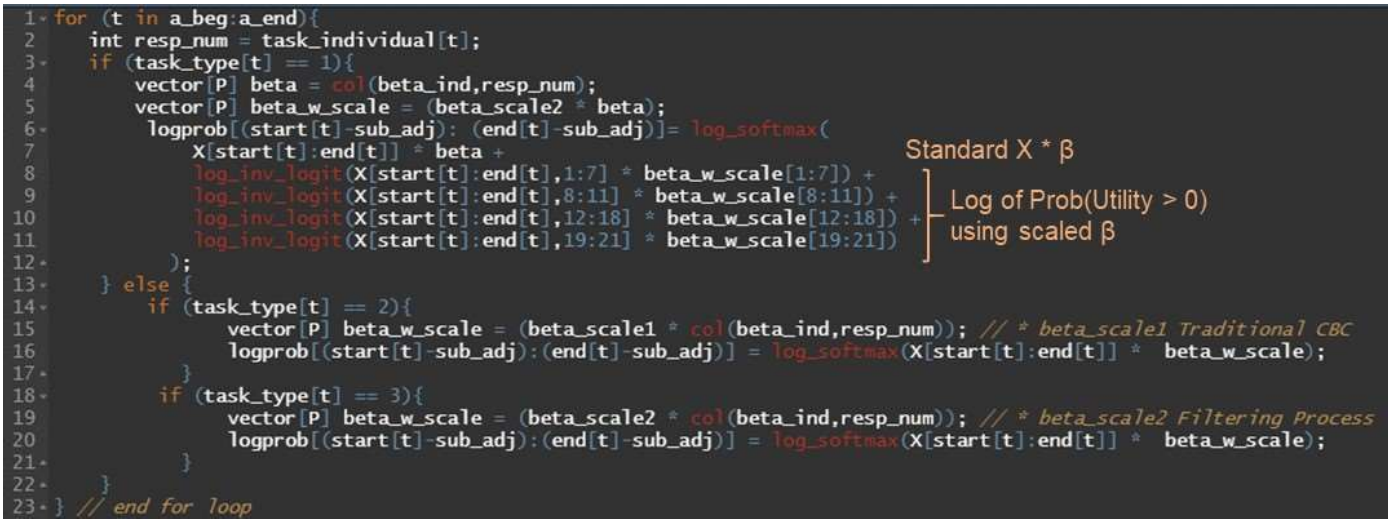
We had similar tasks for brand, data, and contract. In total we added 21 tasks for each respondent, corresponding to the possible levels that could be filtered upon. Since we are comparing each level of an attribute to a 0 vector, we indicator coded each level.

It would be erroneous to assume that a specific level is 0 (dummy coding) or that the sum is 0 (effects coding) when we are assuming the threshold for filtering is 0. For attributes that are not modeled as part of filtering process (like price) we coded in standard ways with dummy or effects coding. Since we had 21 supplemental tasks, and only 6 filter CBC tasks, we weighted supplemental tasks by 6/21.

In addition to partial profile coding the filtering process, we also need to include those probabilities in the filter CBC tasks. For each of the 100 alternatives, some of the alternatives are more likely to be screened out than others. So, the filter CBC tasks have two components to the utility. The first component is the standard utility. But then we also add the log of the probability from conjunctive screening on the 4 attributes:

$$\begin{aligned} \text{Ualt} = \; & X\beta + \\ & \ln(\text{Prob}(U_{a1}>0)) + \\ & \ln(\text{Prob}(U_{a2}>0)) + \\ & \ln(\text{Prob}(U_{a3}>0)) + \\ & \ln(\text{Prob}(U_{a4}>0)) + \\ & \varepsilon \sim \text{Gumbel} \end{aligned}$$

It is important to note that we are estimating the 3 different parts of the model simultaneously. In our Stan code we assign each respondent task a type: 1,2, or 3. Type 1 is filter CBC (with 100 alternatives), type 2 is the traditional CBC, and type 3 is the supplemental partial profile tasks based on the filtering process. The custom likelihood function loops through each task, and based on task type (task_type[t]) it does one of three computations:



```
for (t in a_beg:a_end){
   int resp_num = task_individual[t];
   if (task_type[t] == 1){
       vector[P] beta = col(beta_ind,resp_num);
       vector[P] beta_w_scale = (beta_scale2 * beta);
        logprob[(start[t]-sub_adj): (end[t]-sub_adj)]= log_softmax(
           X[start[t]:end[t]] * beta +
           log_inv_logit(X[start[t]:end[t],1:7] * beta_w_scale[1:7]) +
           log_inv_logit(X[start[t]:end[t],8:11] * beta_w_scale[8:11]) +
           log_inv_logit(X[start[t]:end[t],12:18] * beta_w_scale[12:18]) +
           log_inv_logit(X[start[t]:end[t],19:21] * beta_w_scale[19:21])
         );
    } else {
        if (task_type[t] == 2){
            vector[P] beta_w_scale = (beta_scale1 * col(beta_ind,resp_num)); // * beta_scale1 Traditional CBC
            logprob[(start[t]-sub_adj):(end[t]-sub_adj)] = log_softmax(X[start[t]:end[t]] *  beta_w_scale);
        }
         if (task_type[t] == 3){
            vector[P] beta_w_scale = (beta_scale2 * col(beta_ind,resp_num)); // * beta_scale2 Filtering Process
            logprob[(start[t]-sub_adj):(end[t]-sub_adj)] = log_softmax(X[start[t]:end[t]] *  beta_w_scale);
         }
    }
} // end for loop
```

Standard X * β

Log of Prob(Utility > 0) using scaled β

As before, we have separate scale factors for the traditional CBC (beta_scale1) and the partial profile filtering tasks (beat_scale2).

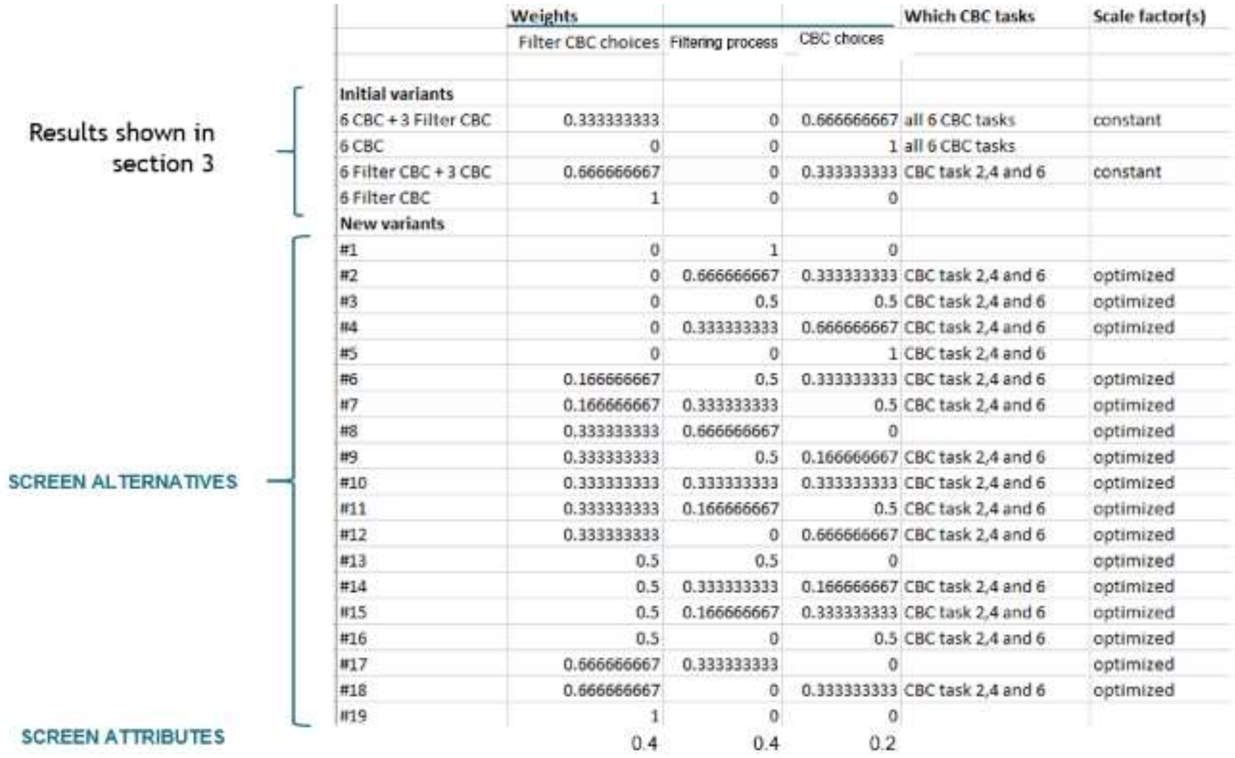
4.3 Comparative Summary and a Few Other Considerations

For screening alternatives, we used a custom Gibbs sampler in R, while we use Stan's sampler for screening attribute levels. As an aside, Lattery believed that 21 price slopes were too many. He prefers to determine the number of price slopes by testing unconstrained aggregate models. With 21 price slopes, an unconstrained aggregate MNL model produces several positive slopes (when we want them to be negative). Using 12 price slopes, the unconstrained aggregate MNL model results in all price slopes <= -.3. A summary of the differences between the two methods is shown in the table below:

Method 1	Method2
Filter = Screen Alternatives	Filter = Screen Attributes
Kees	Kevin
Gibbs Sampler in R	Stan (NUTS sampler)
21 Price Slopes	12 Price Slopes
Weighted 3 Formats in many different ways	Filter Process = Filter Tasks = 6 Traditional CBC = 3

In the table below we show the many different ways the screen alternatives model was weighted.

Exhibit 5: Overview of All the Variants that We Have Tested



	Weights			Which CBC tasks	Scale factor(s)
	Filter CBC choices	Filtering process	CBC choices		
Initial variants					
6 CBC + 3 Filter CBC	0.333333333	0	0.666666667	all 6 CBC tasks	constant
6 CBC	0	0	1	all 6 CBC tasks	
6 Filter CBC + 3 CBC	0.666666667	0	0.333333333	CBC task 2,4 and 6	constant
6 Filter CBC	1	0	0		
New variants					
#1	0	1	0		
#2	0	0.666666667	0.333333333	CBC task 2,4 and 6	optimized
#3	0	0.5	0.5	CBC task 2,4 and 6	optimized
#4	0	0.333333333	0.666666667	CBC task 2,4 and 6	optimized
#5	0	0	1	CBC task 2,4 and 6	
#6	0.166666667	0.5	0.333333333	CBC task 2,4 and 6	optimized
#7	0.166666667	0.333333333	0.5	CBC task 2,4 and 6	optimized
#8	0.333333333	0.666666667	0		optimized
#9	0.333333333	0.5	0.166666667	CBC task 2,4 and 6	optimized
#10	0.333333333	0.333333333	0.333333333	CBC task 2,4 and 6	optimized
#11	0.333333333	0.166666667	0.5	CBC task 2,4 and 6	optimized
#12	0.333333333	0	0.666666667	CBC task 2,4 and 6	optimized
#13	0.5	0.5	0		optimized
#14	0.5	0.333333333	0.166666667	CBC task 2,4 and 6	optimized
#15	0.5	0.166666667	0.333333333	CBC task 2,4 and 6	optimized
#16	0.5	0	0.5	CBC task 2,4 and 6	optimized
#17	0.666666667	0.333333333	0		optimized
#18	0.666666667	0	0.333333333	CBC task 2,4 and 6	optimized
#19	1	0	0		
	0.4	0.4	0.2		

Results shown in section 3

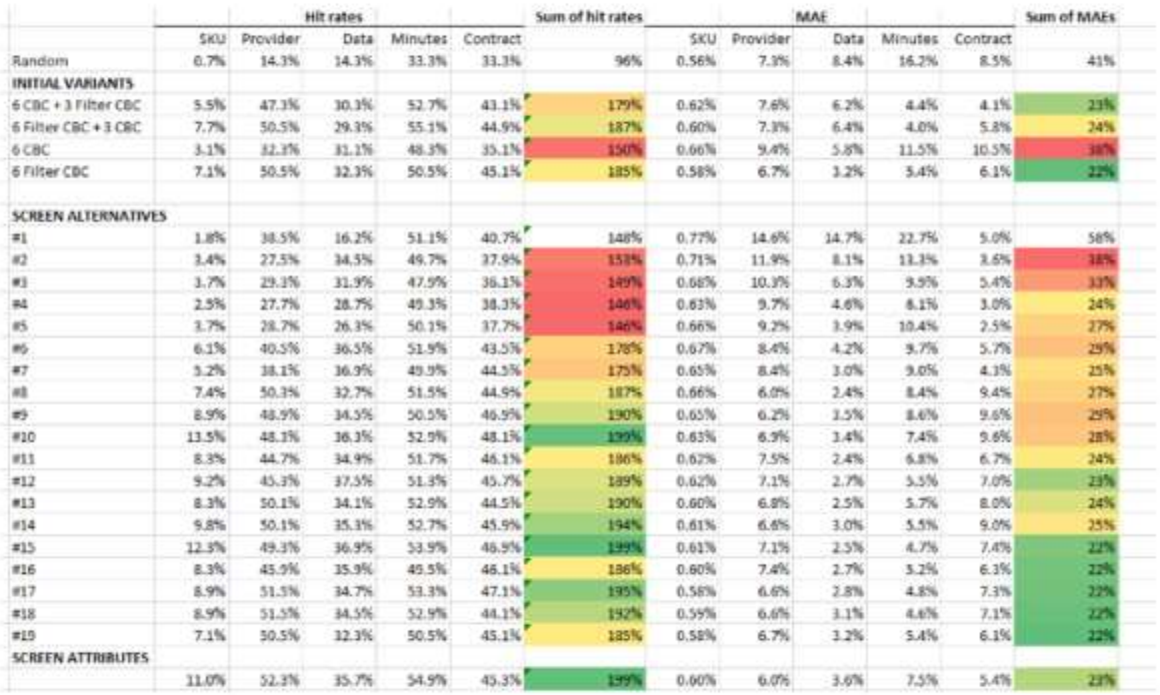
SCREEN ALTERNATIVES

SCREEN ATTRIBUTES

5. Results for Alternate Estimation Methods for Filter CBC

Below in Exhibit 6 the results of everything we tried are listed. We are also showing "random prediction" as a null-level benchmark. Every variant we tested performs better than a random prediction, except variant #1 which is *solely* based on the filtering process data and performs worse on MAE. This variant predicted the same SKU for most respondents, so that was unstable. But it is not surprising, because the filtering process data only is not very informative about the final choices. Other than that, hit rates doubled compared to random in most variants and MAEs halved compared to random in most variants.

Exhibit 6: Overview of the results of all variants that we have tested. For Hit Rate and Mean Absolute Error we took the real-life purchase of respondents as the benchmark.



	Hit rates					Sum of hit rates	MAE					Sum of MAEs
	SKU	Provider	Data	Minutes	Contract		SKU	Provider	Data	Minutes	Contract	
Random	0.7%	14.3%	14.3%	33.3%	33.3%	96%	0.56%	7.3%	8.4%	16.2%	8.5%	41%
INITIAL VARIANTS												
6 CBC + 3 Filter CBC	5.5%	47.3%	30.3%	52.7%	43.1%	179%	0.62%	7.6%	6.2%	4.4%	4.1%	23%
6 Filter CBC + 3 CBC	7.7%	50.5%	29.3%	55.1%	44.9%	187%	0.60%	7.3%	6.4%	4.0%	5.8%	24%
6 CBC	3.1%	32.3%	31.1%	48.3%	35.1%	150%	0.66%	9.4%	5.8%	11.5%	10.5%	38%
6 Filter CBC	7.1%	50.5%	32.3%	50.5%	45.1%	185%	0.58%	6.7%	3.2%	5.4%	6.1%	22%
SCREEN ALTERNATIVES												
#1	1.8%	38.5%	16.2%	51.1%	40.7%	148%	0.77%	14.6%	14.7%	22.7%	5.0%	58%
#2	3.4%	27.5%	34.5%	49.7%	37.9%	153%	0.71%	11.9%	8.1%	11.3%	3.6%	36%
#3	1.7%	29.3%	31.9%	47.9%	36.1%	149%	0.68%	10.3%	6.3%	9.9%	5.4%	33%
#4	2.5%	27.7%	28.7%	49.3%	38.3%	146%	0.63%	9.7%	4.6%	6.1%	3.0%	24%
#5	3.7%	28.7%	26.3%	50.1%	37.7%	146%	0.66%	9.2%	3.9%	10.4%	2.5%	27%
#6	6.1%	40.5%	36.5%	51.9%	43.5%	178%	0.67%	8.4%	4.2%	9.7%	5.7%	29%
#7	5.2%	38.1%	36.9%	49.9%	44.5%	175%	0.65%	8.4%	3.0%	9.0%	4.3%	25%
#8	7.4%	50.3%	32.7%	51.5%	44.9%	187%	0.66%	6.0%	2.4%	8.4%	9.4%	27%
#9	8.9%	48.9%	34.5%	50.5%	46.9%	190%	0.65%	6.2%	3.5%	8.6%	9.6%	29%
#10	13.5%	48.3%	36.3%	52.9%	48.1%	199%	0.63%	6.9%	3.4%	7.4%	9.6%	28%
#11	8.3%	44.7%	34.9%	51.7%	46.1%	186%	0.62%	7.5%	2.4%	6.8%	6.7%	24%
#12	9.2%	45.3%	37.5%	51.3%	45.7%	189%	0.62%	7.1%	2.7%	5.5%	7.0%	23%
#13	8.3%	50.1%	34.1%	52.9%	44.5%	190%	0.60%	6.8%	2.5%	5.7%	8.0%	24%
#14	9.8%	50.1%	35.3%	52.7%	45.9%	194%	0.61%	6.6%	3.0%	5.5%	9.0%	25%
#15	12.3%	49.3%	36.9%	53.9%	46.9%	199%	0.61%	7.1%	2.5%	4.7%	7.4%	22%
#16	8.3%	45.9%	35.9%	49.5%	46.1%	186%	0.60%	7.4%	2.7%	5.2%	6.3%	22%
#17	8.9%	51.5%	34.7%	53.3%	47.1%	195%	0.58%	6.6%	2.8%	4.8%	7.3%	22%
#18	8.9%	51.5%	34.5%	52.9%	44.1%	192%	0.59%	6.6%	3.1%	4.6%	7.1%	22%
#19	7.1%	50.5%	32.3%	50.5%	45.1%	185%	0.58%	6.7%	3.2%	5.4%	6.1%	22%
SCREEN ATTRIBUTES												
	11.0%	52.3%	35.7%	54.9%	45.3%	199%	0.60%	6.0%	3.6%	7.5%	5.4%	23%

Within the 19 "screen alternatives" variants, the trend is clearly that the more weight is given to Filter CBC choices, the better hit rate and MAE we get. Only in the most extreme case when we completely ignore the traditional CBC choices and the filtering process data, the hit rate starts to decrease, but in fact only by a little bit. Combining hit rate and MAE, the best performing variants are #15 and #17. The former has 50% weight for the Filter CBC choices, 17% for the filtering process data and 33% weight for the CBC choices; the latter has 67% weight for the Filter CBC choices and 33% weight for the filtering process. Both these variants do particularly well on predictions at SKU level. Overall, it appears that a high weight for the Filter CBC choices is a must but enhancing it a bit with filtering process data improves the results somewhat further.

The "screen attributes" variant also performs very well, very much comparable with #15 and #17 of the "screen alternatives" variants, while the weight of the Filter CBC choices is only 40%. Since the "screen attributes" rule results in much smaller data files than "screen alternatives," that might be the better way to go forward. Note that the positive result of "screen attributes" might have been caused by a different sampler in HB, although that is not very likely.

6. Conclusions

- Real buying behavior of consumers can be predicted (to a fair amount) by discrete choice modeling.
  - But there is also plenty room to further improve.
- Filter CBC choices on their own (single choice, 6 tasks) already do a great job in predicting correctly.
  - Especially so compared to predictions based on Traditional CBC (single choice, 6 tasks) on their own.
- Enhancing Filter CBC data "a little bit" with other data (the data of the filtering process in particular) improves predictions further.
- There are multiple ways to incorporate the filtering process in the model. Screening by attribute performs very well in comparison with screening by concept and is more efficient in data handling.

Marco Hoogerbrugge

Menno de Jong

Kevin Lattery

Kees van der Wagt

USING MAXDIFF INSTEAD OF CBC?
PROS, CONS, AND RECOMMENDATIONS

MEGAN PEITZ
ABBY LERNER
NUMERIOUS

ABSTRACT

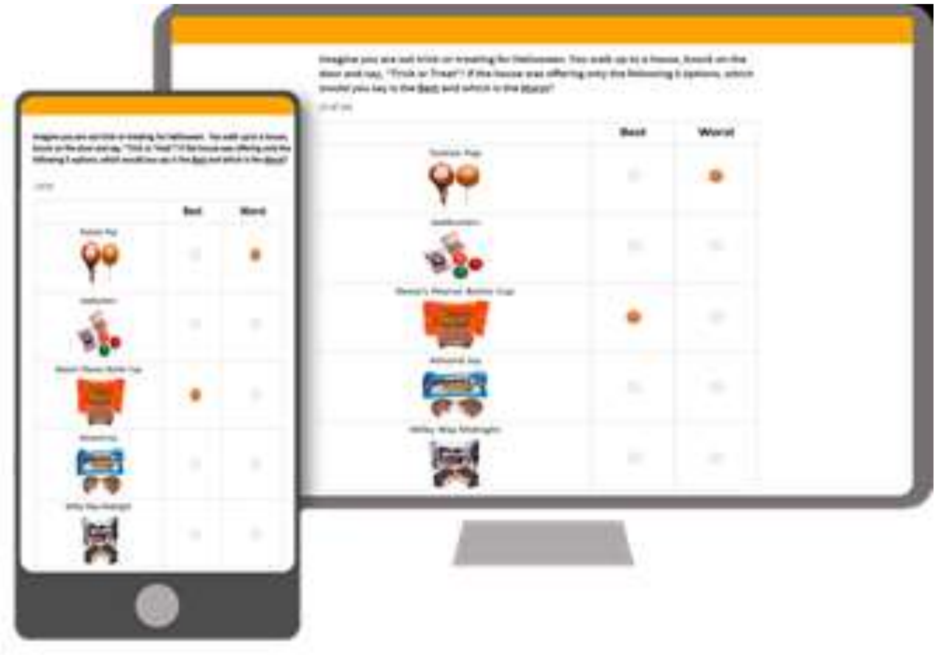
Anchored MaxDiff has given researchers a new set of capabilities when it comes to the analysis of Best-Worst data. Using the Indirect Anchoring approach from Jordan Louviere or the Direct Anchoring approach from Kevin Lattery, we are now able to simulate a "None" alternative that we previously only had in a Choice-Based Conjoint (CBC). But can (and should) we take our smaller CBC designs and turn them into Anchored MaxDiffs? If we do, will we get similar results? This paper sets out to understand the pros and cons of turning a simple CBC into an Anchored MaxDiff and recommendations for implementation. We also test two new approaches to Anchoring MaxDiff data.

INTRODUCTION

MaxDiff versus Choice-Based Conjoint

MaxDiff, otherwise known as Best-Worst Scaling, is an approach for measuring preferences for a list of items. "Items" can include advertising claims, product benefits, product messaging, images, product names, claims, brands, features, packaging options, and more. In a MaxDiff experiment, respondents are typically shown 3-6 items at a time and are asked to indicate which is best and which is worst, or some other relevant alternative. The task is repeated many times, showing a different set of items in each task. The resulting model provides ratio-scaled scores for each item that can be transformed into ranks.

Figure 1.1: Example MaxDiff Task



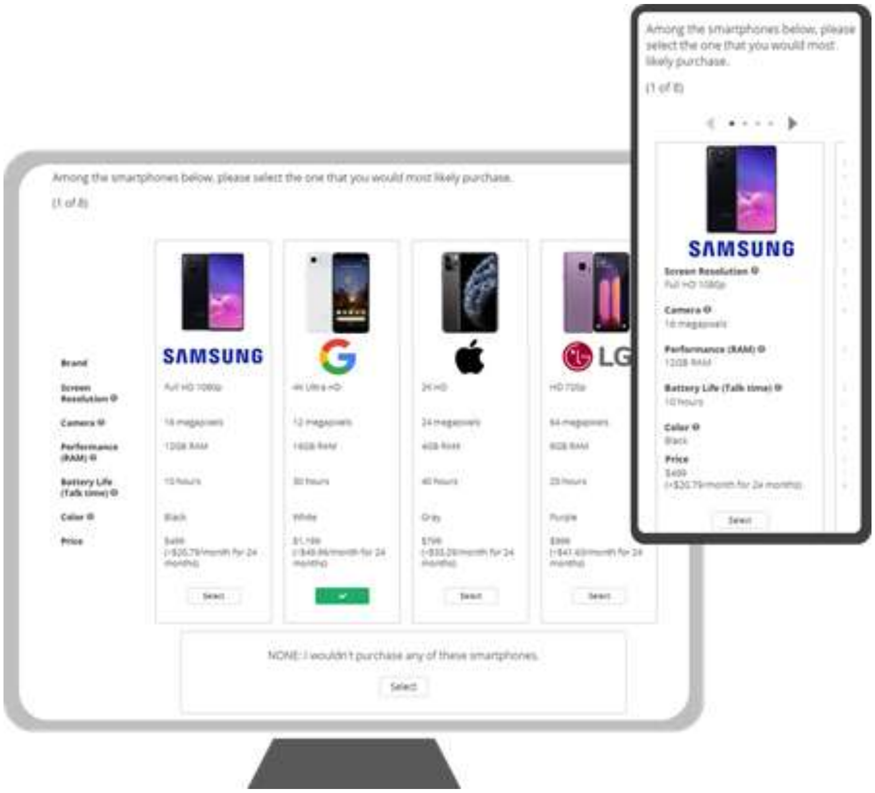
Imagine you are out trick or treating for Halloween. You walk up to a house, knock on the door and say, "Trick or Treat!" If the house was offering only the following 5 options, which would you say is the Best and which is the Worst?

(1 of 10)

	Best	Worst
Tootsie Pop	○	●
Jawbreakers	○	○
Reese's Peanut Butter Cup	●	○
Almond Joy	○	○
Milky Way Midnight	○	○

In a Choice-Based Conjoint (CBC) experiment, we are typically evaluating concepts defined by two or more dimensions. These dimensions are typically referred to as "attributes" (i.e., brand and price). Within each attribute, we define specific levels (i.e., Samsung, Apple, Pixel are levels of the Brand attribute) and an experimental design is used to show respondents different combinations of the levels of the attributes. Respondents are typically shown 3-4 concepts items at a time and are asked to indicate which they are most likely to buy, if any. The task is repeated many times, showing a different set of concepts in each task. The resulting model creates utilities for each level that can be summed across attributes and transformed using the logit rule into probabilities of choice (share of preference scores) for all possible concept combinations. Researchers can then simulate how people would choose given any combination.

Figure 1.2: Example Choice-Based Conjoint Task



Among the smartphones below, please select the one that you would most likely purchase.

(1 of 8)

Brand	SAMSUNG	G		LG
Screen Resolution	Full HD 1080p	4K Ultra HD	2K HD	HD 720p
Camera	16 megapixels	12 megapixels	24 megapixels	64 megapixels
Performance (RAM)	12GB RAM	16GB RAM	4GB RAM	8GB RAM
Battery Life (Talk time)	10 hours	30 hours	40 hours	20 hours
Color	Black	White	Gray	Purple
Price	$499 (≈$20.79/month for 24 months)	$1,199 (≈$49.96/month for 24 months)	$799 (≈$33.29/month for 24 months)	$999 (≈$41.63/month for 24 months)
	Select	✓	Select	Select

NONE: I wouldn't purchase any of these smartphones.

Select

Among the smartphones below, please select the one that you would most likely purchase.

(1 of 8)

SAMSUNG

Screen Resolution
Full HD 1080p

Camera
16 megapixels

Performance (RAM)
12GB RAM

Battery Life (Talk time)
10 hours

Color
Black

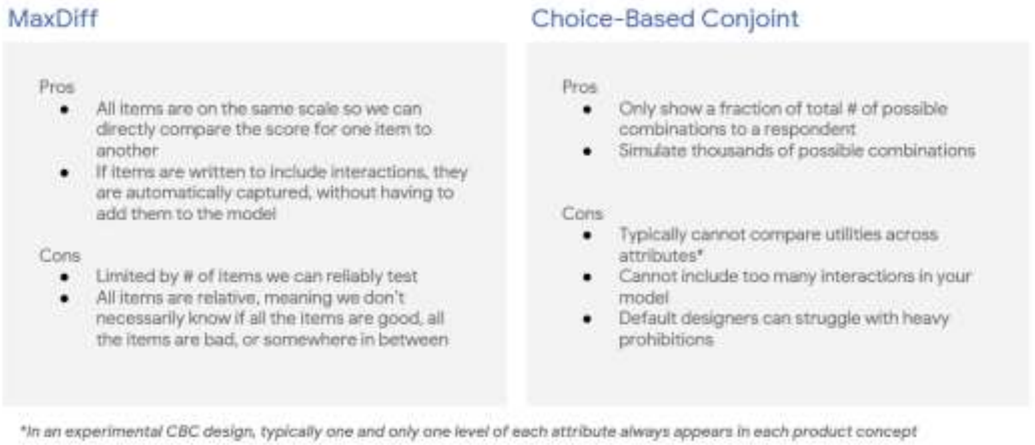
Price
$499
(≈$20.79/month for 24 months)

Select

Both methodologies have their pros and cons when we compare them head-to-head. For example, the results of a MaxDiff place all the items tested on the same scale so that we can directly compare the score for one item to another. Whereas, in a Choice-Based Conjoint, because we tend to show one and only one level of each attribute in each concept, there isn't a common anchoring point across attributes and we typically cannot compare utilities across attributes. This can often be confusing for a client who wants to use simple math to compare the utilities from the brand attribute to the utilities of the price attribute.

Below is a summary of the comparative pros and cons to each method.

Figure 1.3: Pros and Cons of MaxDiff versus CBC



MaxDiff

Pros

- All items are on the same scale so we can directly compare the score for one item to another
- If items are written to include interactions, they are automatically captured, without having to add them to the model

Cons

- Limited by # of items we can reliably test
- All items are relative, meaning we don't necessarily know if all the items are good, all the items are bad, or somewhere in between

Choice-Based Conjoint

Pros

- Only show a fraction of total # of possible combinations to a respondent
- Simulate thousands of possible combinations

Cons

- Typically cannot compare utilities across attributes*
- Cannot include too many interactions in your model
- Default designers can struggle with heavy prohibitions

**In an experimental CBC design, typically one and only one level of each attribute always appears in each product concept*

Which Method Should I Choose?

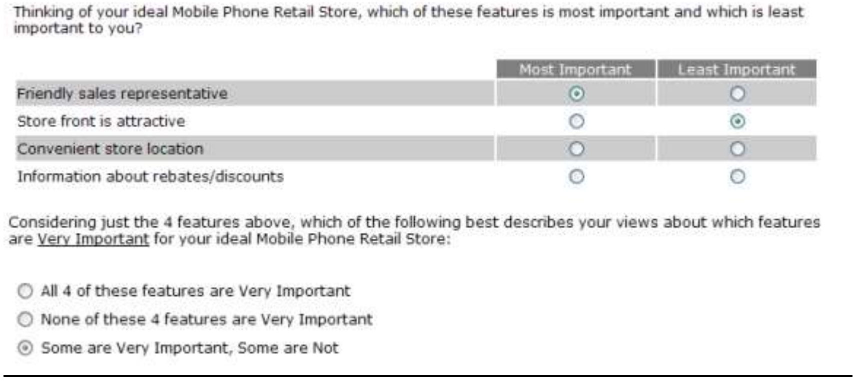
One might consider leveraging MaxDiff instead of CBC for a given business objective, however. For practitioners, it may be a rare occurrence to have a Choice-Based Conjoint (CBC) design small enough to take the exhaustive set of product combinations into a manageable MaxDiff experiment. For example, a CBC with 3 attributes, each with 3 levels, has 3^3 or 27 total combinations. If we want to turn this into a MaxDiff experiment and we follow the traditional recommendations, showing each of the 27 total combinations at least 3 times per respondent (Orme, 2005), we could have a MaxDiff experiment with 17 screens. (27 items * 3 observations per item / 5 items per screen = 16.2 MaxDiff screens).

Odds are though that our design space is bigger than 27 total combinations. Luckily, we have learned from Wirth and Wolfrath (2012) that we can leverage a sparse design, where we show each item to each respondent at least once. For example, with 60 items in the study, 15 sets per respondent and 4 items per set, each item can show 1x per respondent. Or, with 100 items in the study, we can show 20 sets per respondent and 5 items per set. Additionally, Chrzan and Peitz (2019) showed that with large attribute sets, an HB-MNL model that shows each item at least 1x per respondent is comparable to a model with 3x views of each item per respondent (although the more views the better).

Now, assuming we have a list of attributes and levels that is manageable with a MaxDiff design, we can implement an “anchor” within the MaxDiff line of questioning to allow us to model the “None” alternative that is commonplace in a CBC experiment.

There are two anchoring methods currently available within Sawtooth Software’s Lighthouse Studio: the Dual Response Indirect Method from Jordan Louviere, and the Direct Binary Approach Method from Kevin Lattery. In the Indirect approach (Figure 1.4), we add a question to every MaxDiff screen and ask the respondent if all, some, or none of the items are important (alternatively, if they would actually buy all, some, or none of the items in the subset shown).

Figure 1.4: Dual Response Indirect Anchoring Method Example



Thinking of your ideal Mobile Phone Retail Store, which of these features is most important and which is least important to you?

	Most Important	Least Important
Friendly sales representative	◉	○
Store front is attractive	○	◉
Convenient store location	○	○
Information about rebates/discounts	○	○

Considering just the 4 features above, which of the following best describes your views about which features are Very Important for your ideal Mobile Phone Retail Store:

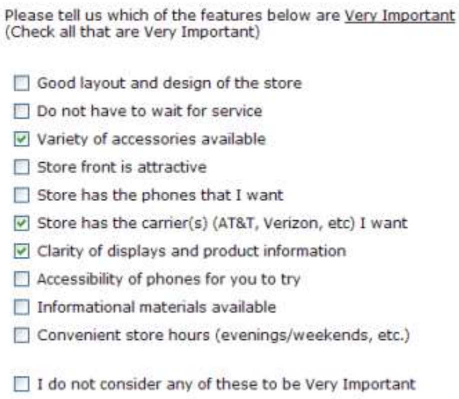
○ All 4 of these features are Very Important

○ None of these 4 features are Very Important

◉ Some are Very Important, Some are Not

In the Direct approach (Figure 1.5), we add one question, either before or after the MaxDiff exercise, and it can either be a select all (i.e., multi-select) or a ratings question, and you can show the entire list of items or a subset.

Figure 1.5: Direct Anchoring Method Example Question



Please tell us which of the features below are Very Important for your ideal Mobile Phone Retail Store?
(Check all that are Very Important)

- □ Good layout and design of the store
- □ Do not have to wait for service
- ☑ Variety of accessories available
- □ Store front is attractive
- □ Store has the phones that I want
- ☑ Store has the carrier(s) (AT&T, Verizon, etc) I want
- ☑ Clarity of displays and product information
- □ Accessibility of phones for you to try
- □ Informational materials available
- □ Convenient store hours (evenings/weekends, etc.)

□ I do not consider any of these to be Very Important

Assuming the researcher has a list of attributes and levels whose exhaustive profile combinations can be managed with a Sparse MaxDiff design, the next question could be whether to use CBC or MaxDiff to design and analyze the data.

And because Sparse MaxDiff did better than Express in the Chrzan and Peitz (2019) work, the hypothesis of the authors was that one observation of every item in MaxDiff could prove greater than CBC where only a fraction of the total combinations of items are shown.

And, if we run a MNL Latent Class analysis on our MaxDiff data, where the items are essentially the entire product combination, we can get clusters based on the entire product preference, not just individual feature preference.

STUDY DESIGN

Attributes and Levels

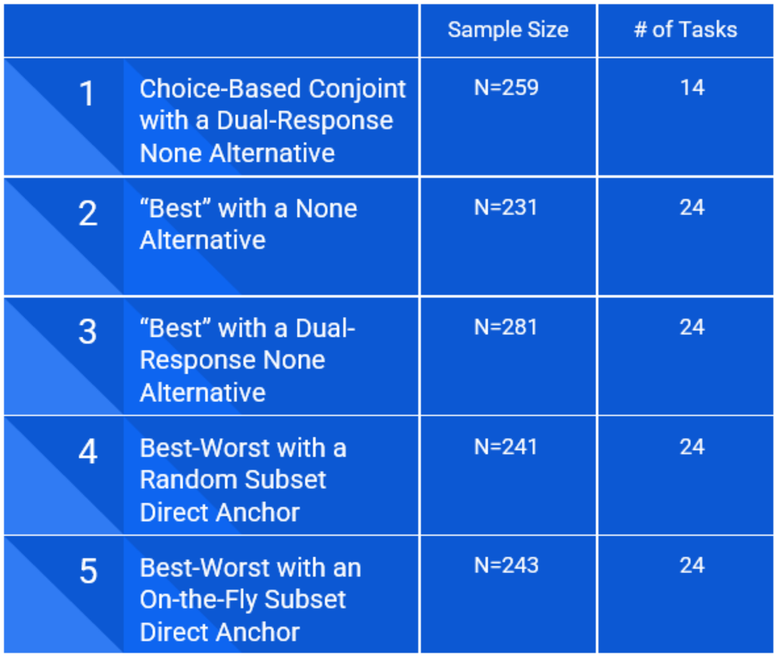
This research uses smart displays as the product to be evaluated. The set of attributes and levels tested is in Table 2.1. We employ an alternative-specific design where the price attribute shown is conditional upon the size of the device. In addition, there are different levels of Voice Assistant for the different brands tested and the Camera attribute is conditional upon the size of the device being at least 10" or 15" (Note—a camera is not available on the 7" device).

Table 2.1: Attributes and Levels Tested

ATTRIBUTE	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
Size	7"	10"	15"	
Brand	Amazon	Apple	Facebook	Google
Voice Assistant	No Voice Assistant	Voice Assistant		
Voice Assistant ***(IF Facebook)***	No Voice Assistant	Works with Amazon Alexa	Works with Google Assistant	Works with Amazon Alexa & Google Assistant
Camera for Video Calling ***(IF 10" or 15")***	No Camera	Includes Camera & Video Calling		
Price of 7"	$89	$99	$129	
Price of 10"	$179	$199	$229	
Price of 15"	$279	$299	$329	

With these alternative-specific relationships in place, there are only 120 total combinations of attributes*levels. From there, each respondent was assigned to one of the following cells.

Figure 2.2: Sample Cells



		Sample Size	# of Tasks
1	Choice-Based Conjoint with a Dual-Response None Alternative	N=259	14
2	"Best" with a None Alternative	N=231	24
3	"Best" with a Dual-Response None Alternative	N=281	24
4	Best-Worst with a Random Subset Direct Anchor	N=241	24
5	Best-Worst with an On-the-Fly Subset Direct Anchor	N=243	24

There is one CBC cell, two "Best" only cells, and two "Best-Worst" cells. The primary difference is that the CBC cell only shows each respondent ~1/2 of the total product combinations. Not to mention, although there are 5 concepts per task, the default CBC designer in Sawtooth Software's Lighthouse Studio could show duplicate concepts as the goal is obtain high efficiency for main effects and first-order interactions. Therefore, it is not required to show all unique concepts. In this case, for example, even though we show 70 concepts (5 concepts*14 tasks), only about 60 concepts are unique on average across all the versions (100). In the Best and Best-Worst cells, all 120 concepts will be shown once to each respondent.

Figure 2.3: Cell 1—CBC with Dual Response None

CBC with DR None

5 concepts per screen

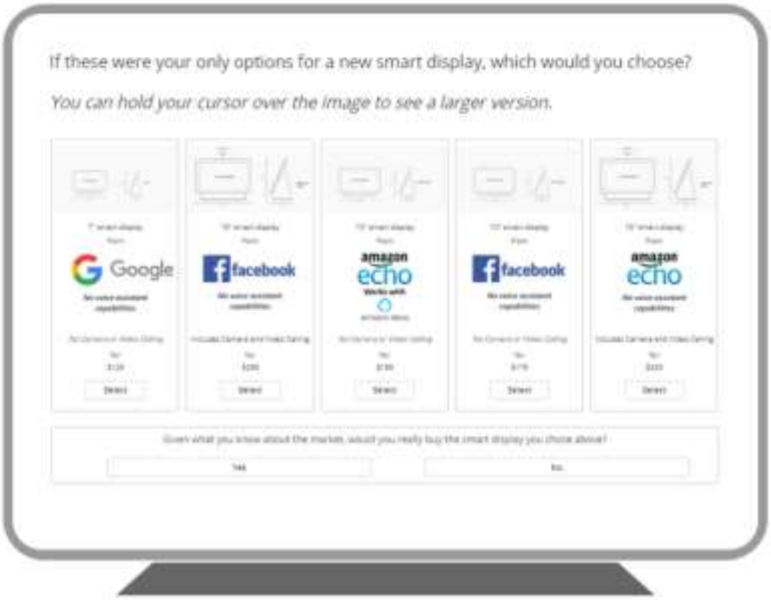
Dual-response none alternative

Sawtooth's CBC Balanced Overlap design with 100 versions

14 total screens

> *Note - we did attempt a "smaller CBC design" with 10 tasks but the default designers returned unestimatable designs*

Respondents only see ½ of the total 120 product combinations



If these were your only options for a new smart display, which would you choose?

You can hold your cursor over the image to see a larger version.

Figure 2.4: Cell 2—"Best" with a None Alternative

Best with Traditional None

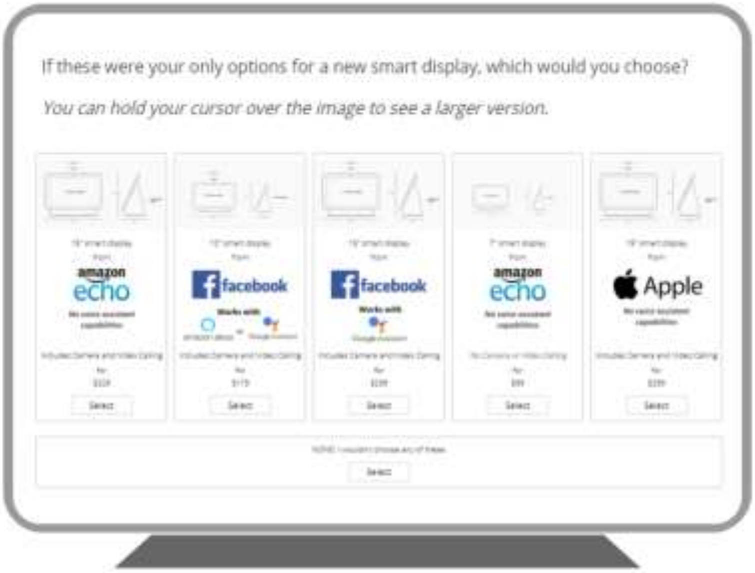
5 items per screen

None alternative on every screen

Ask "Best" only

24 screens so that each of the 120 "items" is shown ~1x

Similar to "Status-Quo" MaxDiff (Chrzan and Lee) except the "None" is on <u>every</u> screen



If these were your only options for a new smart display, which would you choose?

You can hold your cursor over the image to see a larger version.

Figure 2.5: Cell 3—"Best" with a Dual Response None Alternative

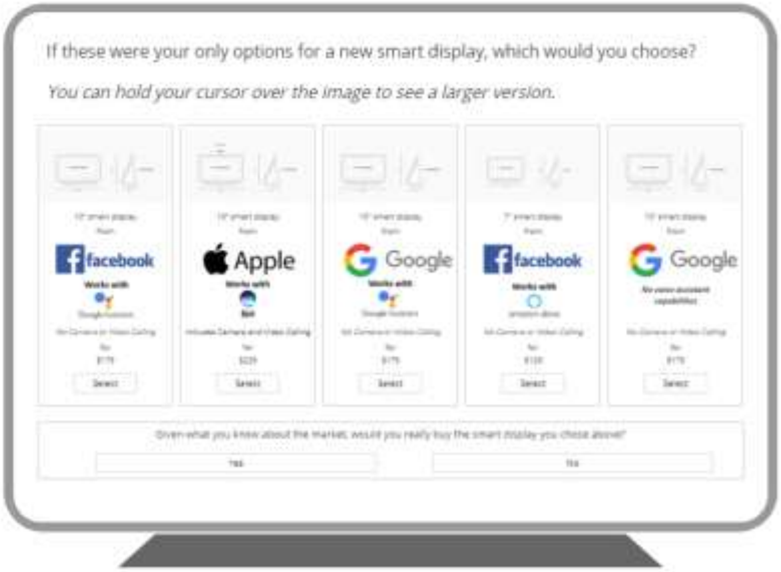
Best with Dual Response None

5 items per screen

Dual Response None-alternative on every screen

Ask "Best" only

24 screens so that each of the 120 "items" is shown ~1x



If these were your only options for a new smart display, which would you choose?

You can hold your cursor over the image to see a larger version.

Given what you know about the market, would you really buy the smart display you chose above?

Figure 2.6: Cell 4—"Best-Worst" with a Random Subset Direct Anchor

Best-Worst with Random Subset Direct Binary Anchor

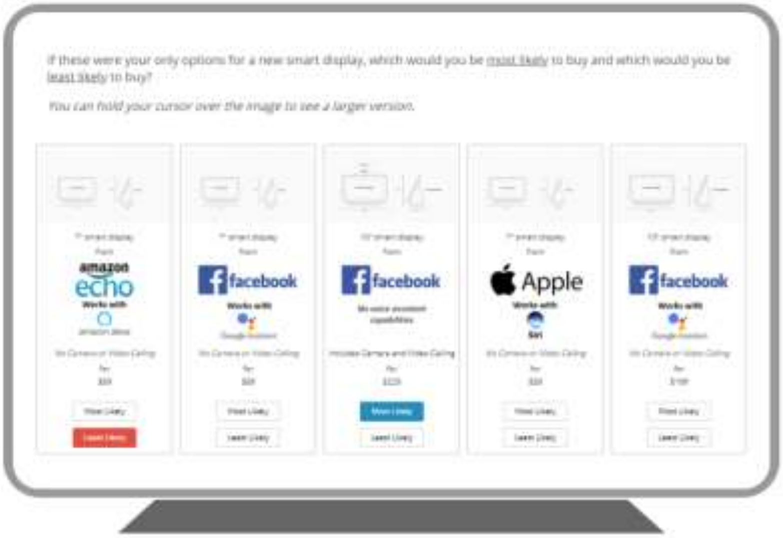
5 items per screen

Ask "Best" and "Worst"

24 screens so that each of the 120 "items" is shown ~1x

Anchor Question - *Now, which of these smart displays, if any, would you actually purchase?*

Select 20 out of 120 random items to be shown in Anchor



If these were your only options for a new smart display, which would you be most likely to buy and which would you be least likely to buy?

You can hold your cursor over the image to see a larger version.

Figure 2.7: Cell 5—"Best-Worst" with an On-the-Fly Subset Direct Anchor

Best-Worst with On-The-Fly Subset Direct Binary Anchor

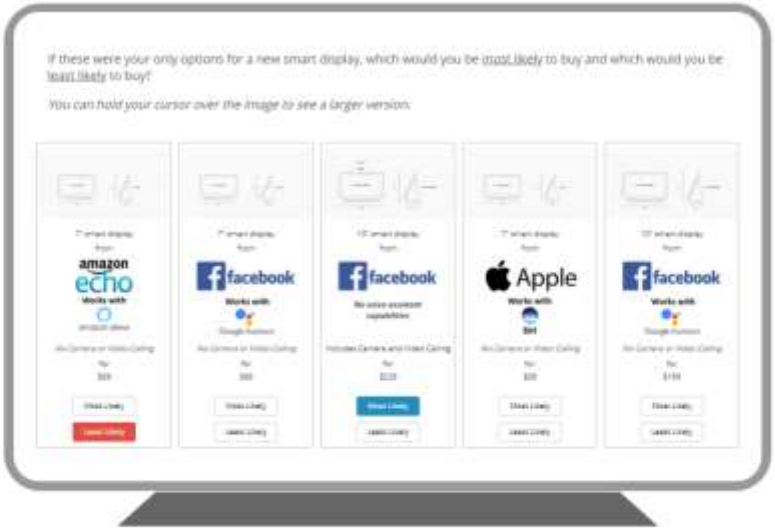
5 items per screen

Ask "Best" and "Worst"

24 screens so that each of the 120 "items" is shown ~1x

Anchor Question - ***Now, which of these smart displays, if any, would you actually purchase?***

Show 20 items based on Sawtooth's "On-the-Fly" logit estimation of ranks 1, 6, 12, 18… , 108, 114, 120



If these were your only options for a new smart display, which would you be most likely to buy and which would you be least likely to buy?

You can hold your cursor over the image to see a larger version.

Holdout Tasks

In a "Holdout" task, the attribute level combinations are fixed—held the same—across respondents. These fixed choice tasks are held out from utility estimation and the part-worth utilities estimated from the responses to the experimentally designed tasks are used to predict respondents' choices to the holdout tasks. The goal is that the predictions closely resemble the answers to the fixed holdout tasks. Chrzan (2015) suggest that at least 5 holdout tasks, if not more, are needed to be confident in these conclusions. This study includes 5 CBC with Dual Response None holdout tasks. We would expect the CBC cell (cell 1) and Best + DR Cell (cell 3) to perform best in the 5 DR holdouts because they most closely resemble the holdout design.

In addition, since MaxDiff results give you a stack rank of preference for all items, we also included 5 ranking holdout questions. In these questions, respondents were asked to rank 5 products, where a rank of 1 was the product they were most likely to buy. The goal is that the respondents' model predictions match the actual rank order of the products.

The Models

In all 5 cells, we created a hierarchical Bayesian (HB) model with prior variance of 1 and 5 degrees of freedom and used point estimates (the default Sawtooth Software settings) after 50K burn-ins and 50K draws. In addition, no constraints or interactions were included in any model.

In order for the scale factor (response error) involved in the different holdout tasks to not affect the share prediction accuracy criterion (MAE), each model's exponent was tuned to minimize the MAE across all holdout tasks. The holdout tasks were not used in estimating the utilities.

THE RESULTS

Across the five cells, we will compare the Mean Absolute Error (MAE), hit rates, and how well the model predicts the "None" category. In addition, price sensitivity curves are compared.

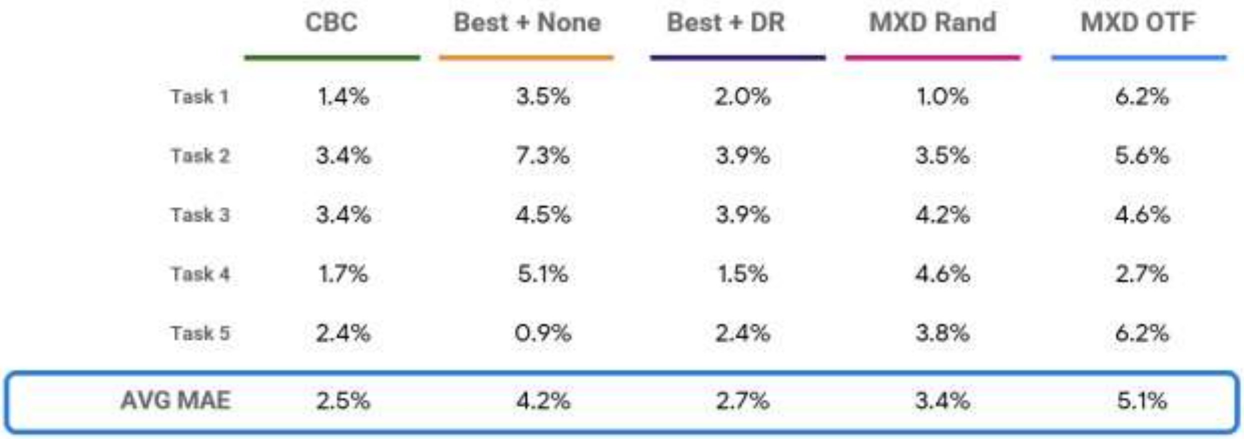
We also believe it is important to have an enjoyable respondent experience to collect high-quality data. Therefore, we will compare the methodologies from the respondent's perspective, examining median time to complete/length of interview (LOI), drop-off percentages, percentage of those who admit to cheating, and respondent evaluations (i.e., easy vs. hard, fun vs. dull).

Comparing Mean Absolute Error

To calculate the Mean Absolute Error (MAE), we build a model to estimate part-worth utilities for each cell. Then we compare the share of preference estimates to actual shares of choice (i.e., frequencies) of all holdout tasks at the aggregate level. We calculate the difference between estimated and actual shares, take the absolute value and average them. The smaller the MAE the better.

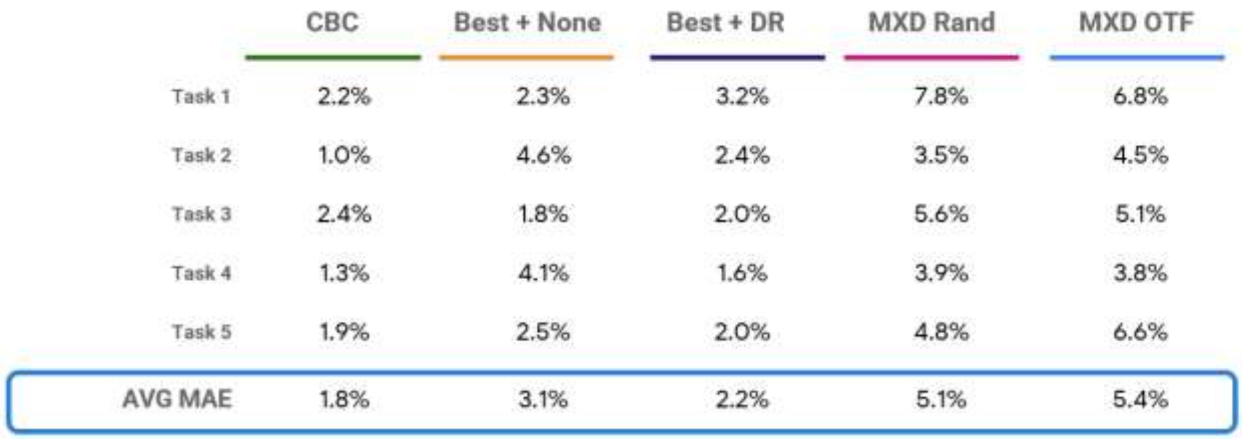
First, for within-sample holdouts, we find that CBC and Best + DR have the lowest MAE (Table 3.1) both when using the model estimates to simulate a scenario including the None alternative and a scenario excluding the None alternative (i.e., only products in the simulation). This finding is expected since the look of the holdout tasks mirrored the look of these two cells compared to the others.

Table 3.1: Within-Sample Mean Absolute Error *Excluding* the None



	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	1.4%	3.5%	2.0%	1.0%	6.2%
Task 2	3.4%	7.3%	3.9%	3.5%	5.6%
Task 3	3.4%	4.5%	3.9%	4.2%	4.6%
Task 4	1.7%	5.1%	1.5%	4.6%	2.7%
Task 5	2.4%	0.9%	2.4%	3.8%	6.2%
AVG MAE	2.5%	4.2%	2.7%	3.4%	5.1%

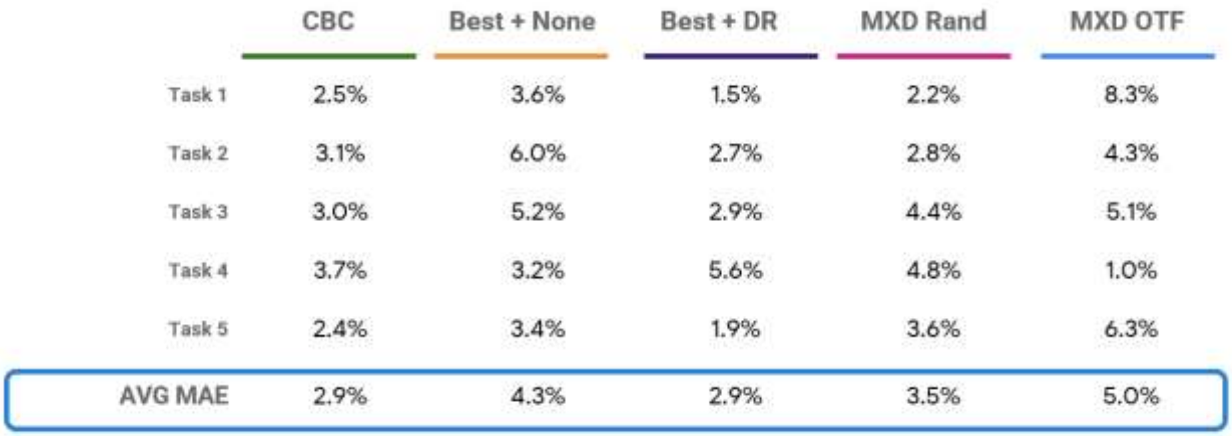
Table 3.2: Within-Sample Mean Absolute Error *Including* the None



	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	2.2%	2.3%	3.2%	7.8%	6.8%
Task 2	1.0%	4.6%	2.4%	3.5%	4.5%
Task 3	2.4%	1.8%	2.0%	5.6%	5.1%
Task 4	1.3%	4.1%	1.6%	3.9%	3.8%
Task 5	1.9%	2.5%	2.0%	4.8%	6.6%
AVG MAE	1.8%	3.1%	2.2%	5.1%	5.4%

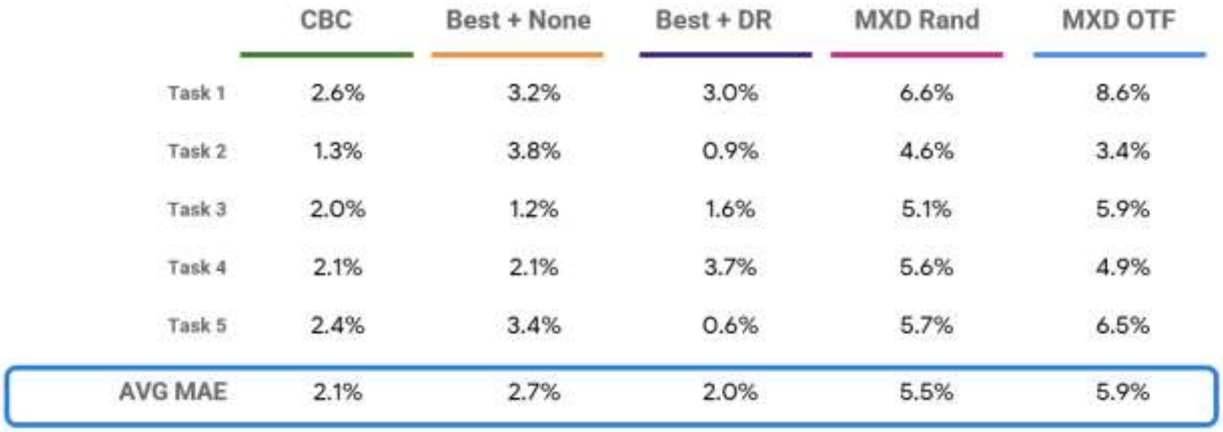
While the previous tables report within-sample holdouts (i.e., comparing the simulated shares to the actual shares using the same respondents), we can also look at out-of-sample data (i.e., comparing the simulated shares from respondents of one cell to the actual shares using respondents from the other four cells). The out-of-sample MAEs are best for the CBC and Best + DR cells. Similarly, when including the None alternative, cells 4 and 5 perform the worst for both within and out-of-sample MAEs (Table 3.2, 3.4).

Table 3.3: Out-of-Sample Mean Absolute Error *Excluding* the None



	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	2.5%	3.6%	1.5%	2.2%	8.3%
Task 2	3.1%	6.0%	2.7%	2.8%	4.3%
Task 3	3.0%	5.2%	2.9%	4.4%	5.1%
Task 4	3.7%	3.2%	5.6%	4.8%	1.0%
Task 5	2.4%	3.4%	1.9%	3.6%	6.3%
AVG MAE	2.9%	4.3%	2.9%	3.5%	5.0%

Table 3.4: Out-of-Sample Mean Absolute Error *Including* the None

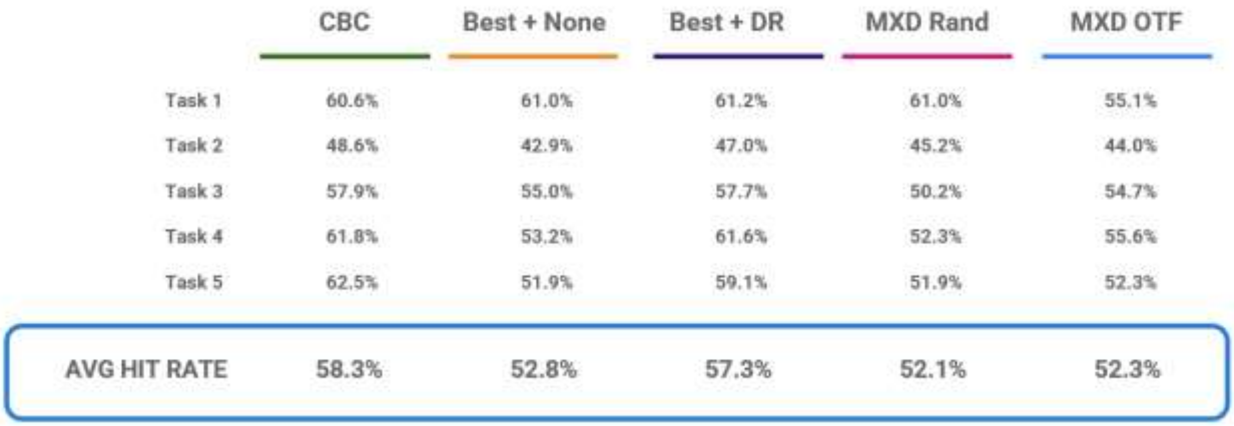


	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	2.6%	3.2%	3.0%	6.6%	8.6%
Task 2	1.3%	3.8%	0.9%	4.6%	3.4%
Task 3	2.0%	1.2%	1.6%	5.1%	5.9%
Task 4	2.1%	2.1%	3.7%	5.6%	4.9%
Task 5	2.4%	3.4%	0.6%	5.7%	6.5%
AVG MAE	2.1%	2.7%	2.0%	5.5%	5.9%

Comparing Hit Rates

In order to calculate the hit rate, we use the same models but compare the simulated share of preference estimates to actual shares of choice (i.e., frequencies) of all holdout tasks at the individual level. If the product chosen in the simulation matches the product chosen in the holdout task for that respondent, then it is counted as a "hit." If it does not match, it is a miss. We then take the sum of hits across all respondents divided by the total correct possible. The higher the hit rate the better.

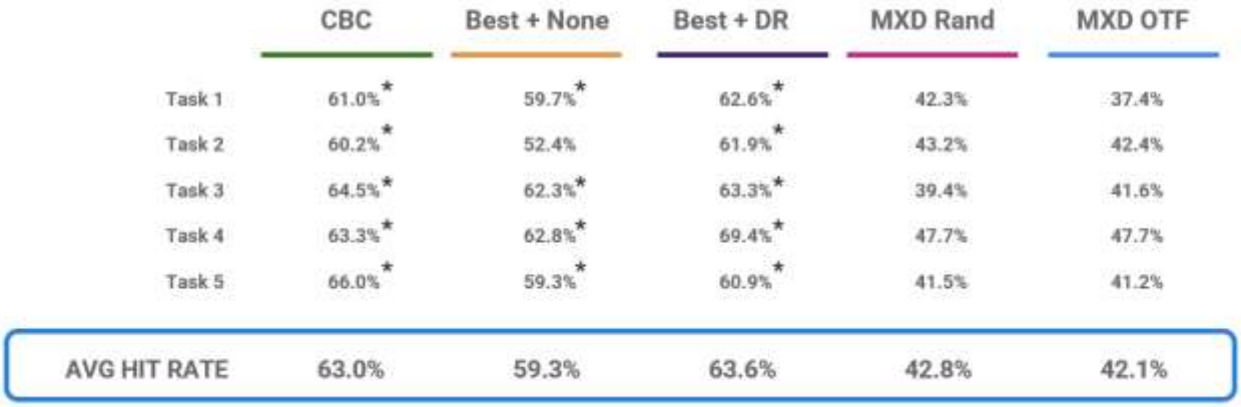
Table 3.5: Hit Rates *Excluding* the None



	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	60.6%	61.0%	61.2%	61.0%	55.1%
Task 2	48.6%	42.9%	47.0%	45.2%	44.0%
Task 3	57.9%	55.0%	57.7%	50.2%	54.7%
Task 4	61.8%	53.2%	61.6%	52.3%	55.6%
Task 5	62.5%	51.9%	59.1%	51.9%	52.3%
AVG HIT RATE	58.3%	52.8%	57.3%	52.1%	52.3%

When including the none, CBC and Best + DR still have the highest average hit rates. The Best + None also separates itself from the Direct Anchoring methods.

Table 3.6: Hit Rates *Including* the None



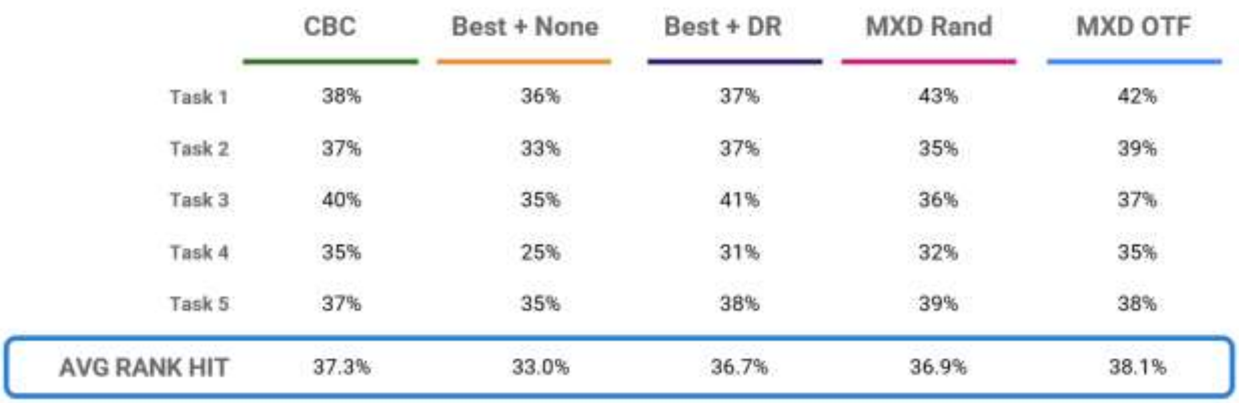
	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	61.0%*	59.7%*	62.6%*	42.3%	37.4%
Task 2	60.2%*	52.4%	61.9%*	43.2%	42.4%
Task 3	64.5%*	62.3%*	63.3%*	39.4%	41.6%
Task 4	63.3%*	62.8%*	69.4%*	47.7%	47.7%
Task 5	66.0%*	59.3%*	60.9%*	41.5%	41.2%
AVG HIT RATE	63.0%	59.3%	63.6%	42.8%	42.1%

**Significantly different from MXD Rand & MXD OTF*

Comparing Rank Orders

Similar to the hit rate, we can compare the rank order from estimated utilities for an individual to actual ranks across all the ranking holdout tasks. A "hit" is counted if the simulated rank order matches the actual rank from the holdout task. Then we take the sum of hits across all respondents divided by the total correct possible. The higher the rank hit rate the better. Note—In this calculation, only an exact rank match is counted and there are 5!, or 120 possible ways that 5 items can be ranked.

Table 3.7: Ranking Hit Rates



	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Task 1	38%	36%	37%	43%	42%
Task 2	37%	33%	37%	35%	39%
Task 3	40%	35%	41%	36%	37%
Task 4	35%	25%	31%	32%	35%
Task 5	37%	35%	38%	39%	38%
AVG RANK HIT	37.3%	33.0%	36.7%	36.9%	38.1%

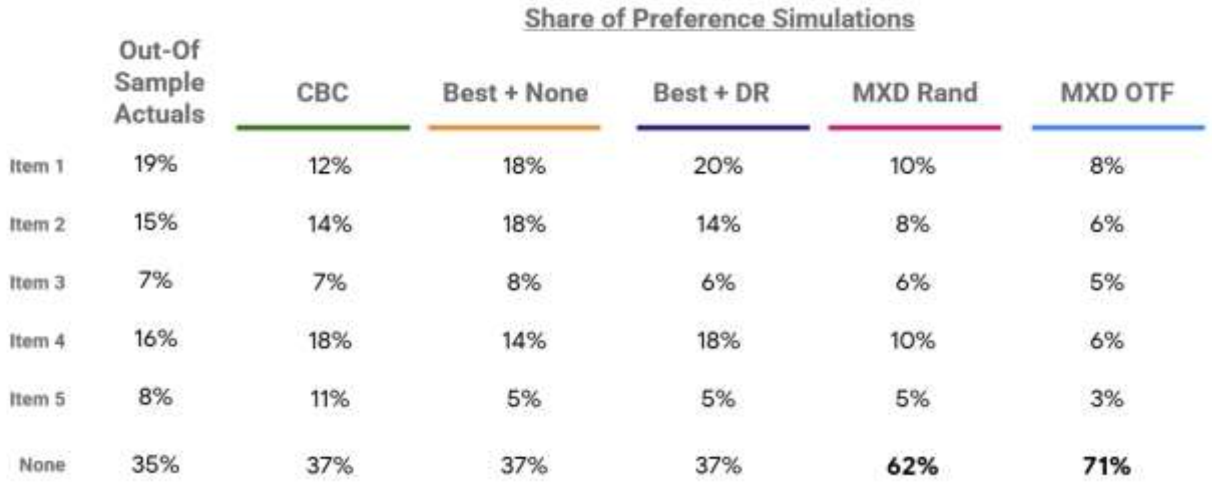
In the ranking hit rates, all methods perform relatively equivalent, except the Best + None cell. In this cell, on average, respondents are selecting none 25% of the time. This perhaps is resulting in too little information at the individual item level in a sparse design to do well predicting some of those middle-ranked items.

Unintended Finding

What is happening with the Direct Anchor?

When examining the results, in each of the direct anchor cells (Cell 4 and 5), when the None is included in the simulation, we are unable to tune the exponent to get an MAE score in parity to the other cells. If we look at the simulation without tuning the exponent (Table 3.5), we can see that the model simulates respondents in Cells 4 and 5 are two times more likely to select "None" as they are in the other cells.

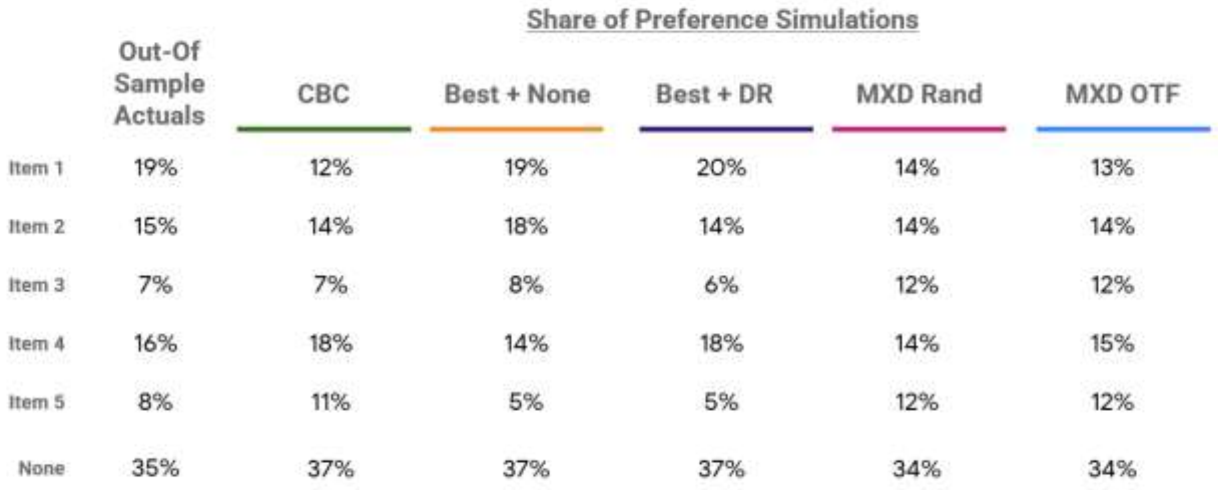
Table 3.8: Raw Simulations from Holdout #3



		Share of Preference Simulations				
	Out-Of Sample Actuals	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Item 1	19%	12%	18%	20%	10%	8%
Item 2	15%	14%	18%	14%	8%	6%
Item 3	7%	7%	8%	6%	6%	5%
Item 4	16%	18%	14%	18%	10%	6%
Item 5	8%	11%	5%	5%	5%	3%
None	35%	37%	37%	37%	**62%**	**71%**

After tuning the exponent (Table 3.9), although the MAEs look better, you can see that the shares are flattened in cells 4 and 5.

Table 3.9: Exponent is Tuned within Each Cell to Minimize MAE



		Share of Preference Simulations				
	Out-Of Sample Actuals	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Item 1	19%	12%	19%	20%	14%	13%
Item 2	15%	14%	18%	14%	14%	14%
Item 3	7%	7%	8%	6%	12%	12%
Item 4	16%	18%	14%	18%	14%	15%
Item 5	8%	11%	5%	5%	12%	12%
None	35%	37%	37%	37%	34%	34%

Our hypotheses around the extreme differences when simulating the None with the Direct Anchor is that we are only showing a subset of all the items in the anchor and thus not getting a complete picture of what respondents are *actually* willing to buy. In addition, because the items are displayed in a list that you vertically scroll through, respondents may have been biased to only choose a few items (as they likely would only ever buy one or two of these devices at a time).

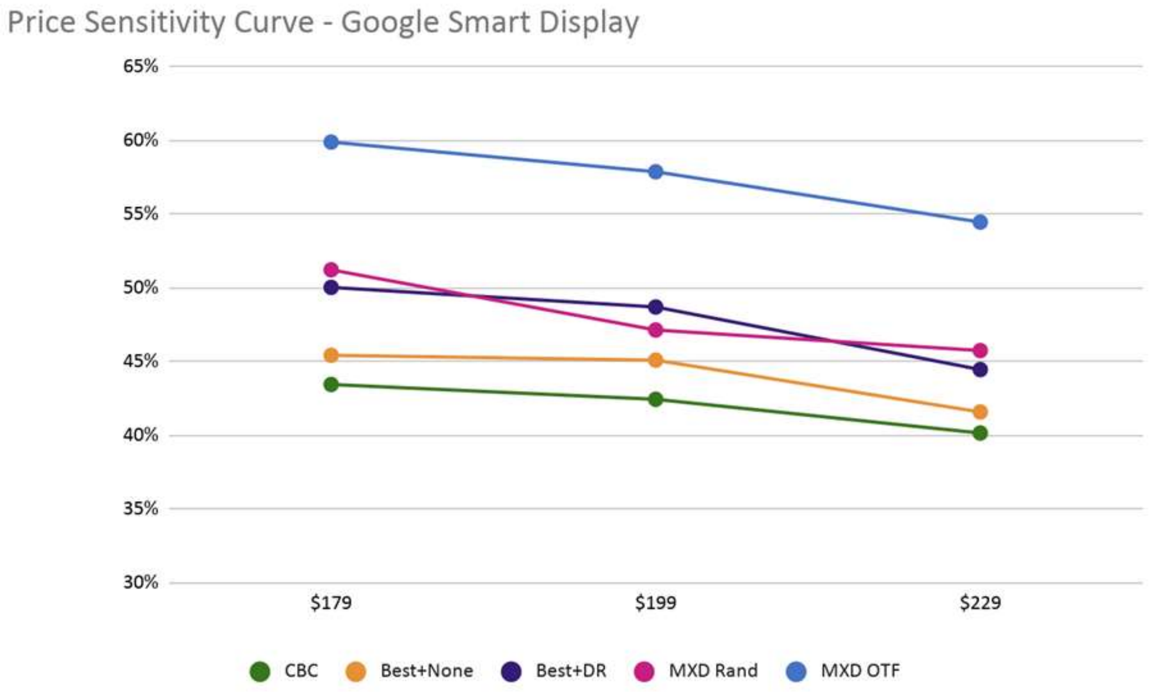
Based on this finding, if using the Direct Anchored MaxDiff for simulations versus the "None," researchers should consider including holdouts for proper tuning of the anchor. In addition, more research should be done around using the Direct Anchor approach with a large list of items (>40), particularly if only showing a subset of the items in the anchor.

Price Sensitivity

Price sensitivity is a primary deliverable of choice research. Therefore, we simulated an Amazon device, a Facebook device and a Google device and graphed the share of preference for the Google device estimated from each model when only changing the price of the Google device.

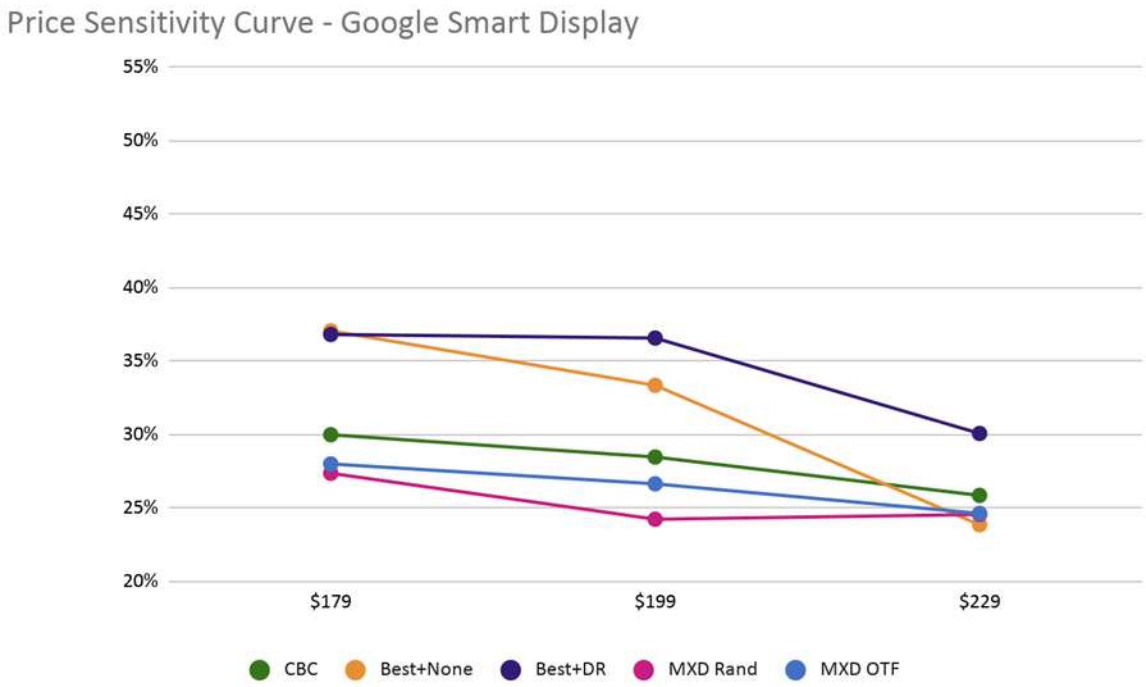
When excluding the None (Figure 4.1), the Best + None most closely aligns with the CBC price sensitivity curve.

Figure 4.1: Price Sensitivity of Google Device *Excluding* the None



When including the None (Figure 4.2), the Best + None and Best + DR pick up more sensitivity to the change in price from $199 to $229. However, this sensitivity is not seen in the CBC cell.

Figure 4.2: Price Sensitivity of Google Device *Including* the None

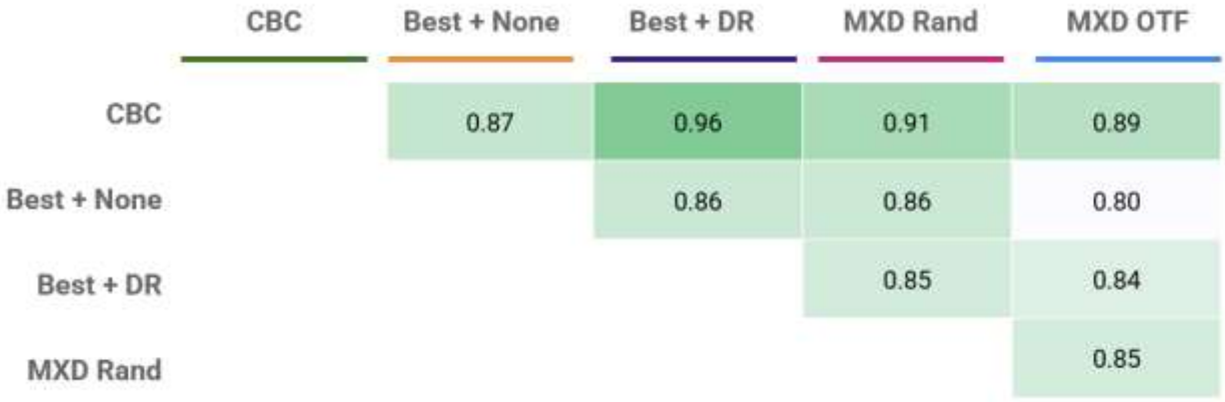


Correlations

Lastly, we can transform the aggregate total utility for each item into a rank across the five cells. Then we can use Spearman's Rank correlation to determine how correlated the rank ordering of the 120 items are across each of the cells.

Overall, the correlations are very strong. With Best + DR and CBC having the strongest correlation (Figure 4.3).

Figure 4.3: Correlation of Rank of 120 Items



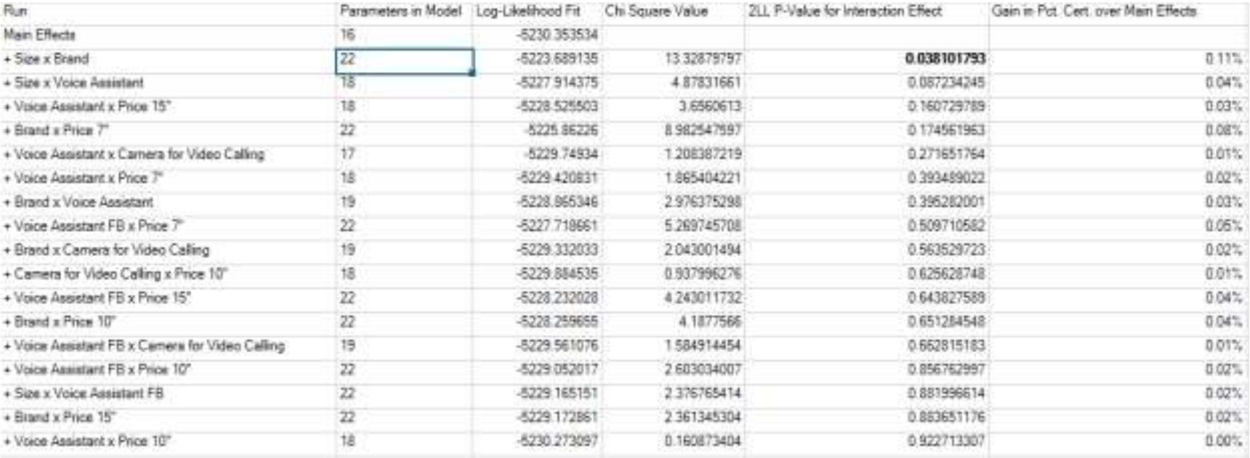
	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
CBC		0.87	0.96	0.91	0.89
Best + None			0.86	0.86	0.80
Best + DR				0.85	0.84
MXD Rand					0.85

Model Interactions

One final issue to address is the perceived benefit of the MaxDiff cells in this paper being able to automatically capture all possible interactions in the model. However, the downside to this approach of a model with 120 parameters is the potential for overfitting.

Sawtooth Software offers an aggregate logit interaction test, or the 2-log likelihood test, that allows the researcher to quickly see if there are any significant interactions to consider including. When running this test on the CBC cell, only one interaction is deemed significant, and even still, the gain in percent certainty over the main effects is minimal (Figure 4.4). Not to mention that even if we did add this interaction, we'd only be up to 22 parameters, a small fraction of the 120 in the MaxDiff cells.

Figure 4.4: Interaction Search Results on CBC Cell



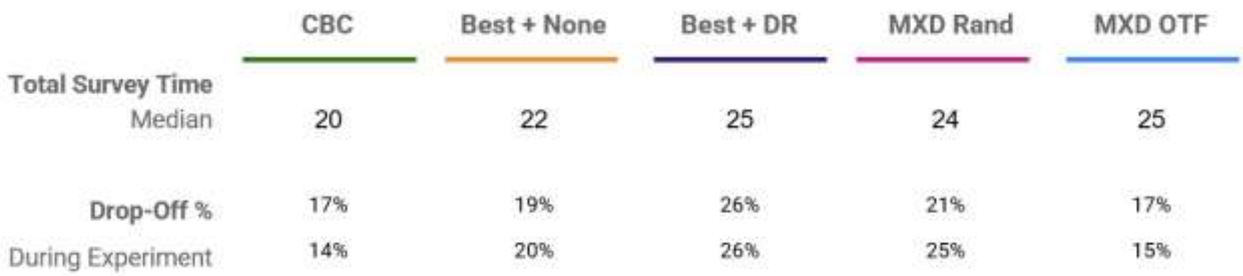
Run	Parameters in Model	Log-Likelihood Fit	Chi Square Value	2LL P-Value for Interaction Effect	Gain in Pct. Cert. over Main Effects
Main Effects	16	-5230.353534			
+ Size x Brand	22	-5223.689135	13.32879797	**0.038101793**	0.11%
+ Size x Voice Assistant	18	-5227.914375	4.87831661	0.087234245	0.04%
+ Voice Assistant x Price 15"	18	-5228.525503	3.6560613	0.160729789	0.03%
+ Brand x Price 7"	22	-5225.86226	8.982547597	0.174561963	0.08%
+ Voice Assistant x Camera for Video Calling	17	-5229.74934	1.208387219	0.271651764	0.01%
+ Voice Assistant x Price 7"	18	-5229.420831	1.865404221	0.393489022	0.02%
+ Brand x Voice Assistant	19	-5228.865346	2.976375298	0.395282001	0.03%
+ Voice Assistant FB x Price 7"	22	-5227.718661	5.269745708	0.509710582	0.05%
+ Brand x Camera for Video Calling	19	-5229.332033	2.043001494	0.563529723	0.02%
+ Camera for Video Calling x Price 10"	18	-5229.884535	0.937996276	0.625628748	0.01%
+ Voice Assistant FB x Price 15"	22	-5228.232028	4.243011732	0.643827589	0.04%
+ Brand x Price 10"	22	-5228.259655	4.1877566	0.651284548	0.04%
+ Voice Assistant FB x Camera for Video Calling	19	-5229.561076	1.584914454	0.662815183	0.01%
+ Voice Assistant FB x Price 10"	22	-5229.052017	2.603034007	0.856762997	0.02%
+ Size x Voice Assistant FB	22	-5229.165151	2.376765414	0.881996614	0.02%
+ Brand x Price 15"	22	-5229.172861	2.361345304	0.883651176	0.02%
+ Voice Assistant x Price 10"	18	-5230.273097	0.160873404	0.922713307	0.00%

Therefore, one can conclude that in this example, a main effects only CBC model outperforms all the exhaustive designs.

Respondent Preference

The CBC cell is the quickest exercise to complete, while the Best + DR cell is the longest (Figure 5.1). Best + DR also has the highest drop off percentage, likely due to its length.

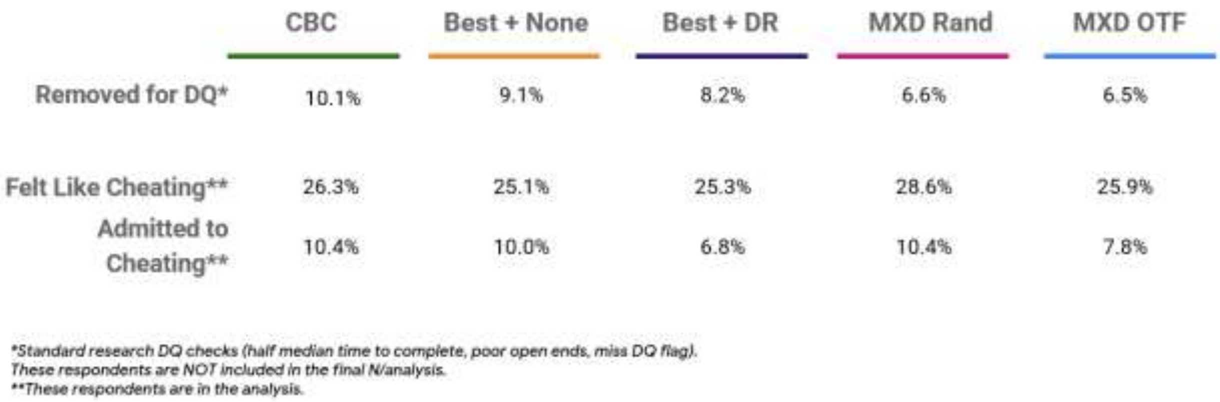
Figure 5.1: Survey Metrics



	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Total Survey Time Median	20	22	25	24	25
Drop-Off %	17%	19%	26%	21%	17%
During Experiment	14%	20%	26%	25%	15%

Throughout the paper, bad data is defined as anyone who admitted to cheating throughout the exercise or who had two flags in the data (speeding, poor RLH, straight lining, etc.). At a glance, the CBC cells seems to have the most bad data, although the differences across cells are not significant (Figure 5.2).

Figure 5.2: Respondent Accuracy

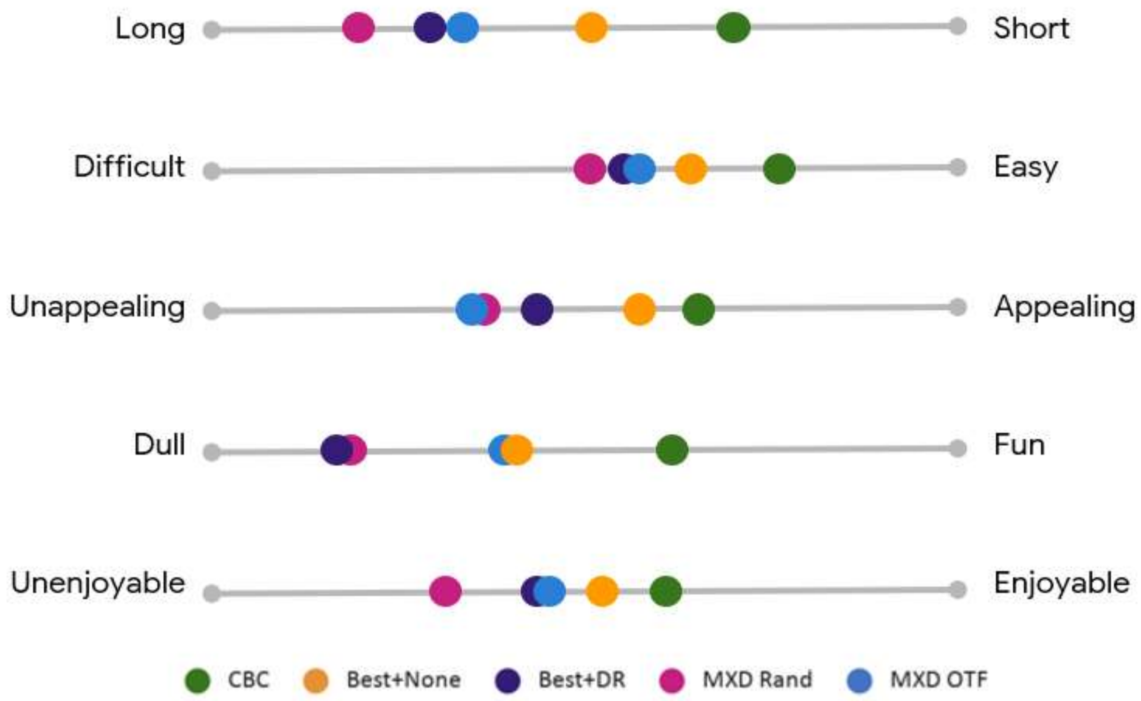


	CBC	Best + None	Best + DR	MXD Rand	MXD OTF
Removed for DQ*	10.1%	9.1%	8.2%	6.6%	6.5%
Felt Like Cheating**	26.3%	25.1%	25.3%	28.6%	25.9%
Admitted to Cheating**	10.4%	10.0%	6.8%	10.4%	7.8%

*Standard research DQ checks (half median time to complete, poor open ends, miss DQ flag).
These respondents are NOT included in the final N/analysis.
**These respondents are in the analysis.

When rating the different methodologies, CBC feels significantly shorter than the other methods and seems to be easier, more appealing, more fun, and more enjoyable (Figure 5.3). The Best + None cell is rated second best by respondents.

Figure 5.3: Respondent Survey Experience Rating



Conclusion

Overall Recommendations

If we distilled all the findings, Best + DR seems to be a viable approach, but CBC is still the best overall, from both a respondent and model perspective. (Again, with the caveat that the holdouts would be expected to favor the CBC-looking approaches.) Therefore, even when the experimental design space can be managed with a sparse MaxDiff design, one should stick to CBC.

Considerations for Future Research

While the Best + Dual Response cell most closely aligns with the CBC results, it is arduous for the respondent. Perhaps the approach can be revisited with a more manageable list of items. In addition, one could consider improving this cell further by asking Best & Worst + DR None, assuming the list is manageable.

The Best + None cell is well liked by respondents, and simulations do well, but we lose too much information by allowing the "None" alternative to appear on every screen. Therefore, it is likely not a viable option when the list of items is long and a Sparse design is necessary.

With a large list of items, only showing a subset in the Direct Anchor (random or OTF) seems to be risky. We are concerned about how severely the exponent needed to be tuned to align with simulations from the other three cells. Therefore, if simulations are your goal, and your list of items is long, stick with a CBC approach.

Megan Peitz

Abby Lerner

References

Chrzan, Keith and Megan Peitz (2019), "Best-Worst Scaling with Many Items," Journal of Choice Modeling, Vol. 30, March 2019, pp 61–72. Accessed at: https://www.sciencedirect.com/science/article/pii/S1755534517301355?via%3Dihub

Lattery, Kevin (2010), "Anchoring Maximum Difference Scaling against a Threshold-Dual Response and Direct Binary Responses," 2010 Sawtooth Software Conference Proceedings. Accessed at: https://www.sawtoothsoftware.com/download/techpap/2010Proceedings.pdf

Orme, Bryan (2005), "Accuracy of HB Estimation in MaxDiff Experiments," Technical Paper available at www.sawtoothsoftware.com

Orme, Bryan (2009), "Anchored Scaling in MaxDiff Using Dual Response," Technical Paper. Accessed at: https://sawtoothsoftware.com/resources/technical-papers/anchored-scaling-in-maxdiff-using-deal-response

Wirth, Ralph, Anette Wolfrath (2012), "Using MaxDiff to Evaluate Very Large Sets of Items," 2012 Sawtooth Software Conference Proceedings, Provo, UT. Accessed at: https://www.sawtoothsoftware.com/download/techpap/2012Proceedings.pdf

Concept Screening and Evaluation Using Survey Based Artificial Markets

John Pemberton
Lacy School of Business, Butler University
Alfred Johnson III
Kelley School of Business, Indiana University
Emily Gasparro
Alyssa MacDonald
Lauren Piskorski
Lacy School of Business, Butler University

Abstract

Screening and evaluation of marketing concepts have evolved over many years. Initially, product concepts were evaluated using traditional research methods such as focus groups and surveys featuring monadic or sequential monadic designs with various Likert scale metrics. More recent techniques for concept screening and evaluation have involved choice-based methods such as Best-Worst Scaling (Louviere, 2015). Building on advances in Prediction markets, concept securities trading games are now also utilized for concept evaluation (Dahan et al., 2011). Various challenges of implementation, reliability, respondent understanding or fatigue present themselves to researchers utilizing these approaches.

The authors describe an approach that utilizes an artificial market in which a research panel responds to sequential monadic exposures of concepts offered and priced within a prediction game. Within the game, "price" represents a value at which a player evaluates the relative market potential of a tested concept. When presented with a purchase opportunity, a player may choose to purchase or sell the concept at the presented price based on their informed perceptions of value across the aggregate market. Points are collected for correct decisions made by the player relative to the final indifference price point identified within the market. Pricing within the game can utilize a learning algorithm to progressively narrow the "offer price" presented to respondents for the tested concept to a precise range where a multivariate statistical technique identifies the final "indifference price." This indifference price is the point at which the aggregate market is equally as likely to buy or sell the product.

The authors believe the implementation of such a market has the potential to be a more reliable predictor than other traditional methodologies. The artificial market will

be tested with a static panel that will also complete MaxDiff (Best-Worst) evaluations of an identical set of tested concepts.

1. Introduction

Successful product launches and marketing campaigns depend on the firm's ability to create products of general need and interest to the consumer, and effective advertising copy to communicate the virtues and benefits of brands or products being promoted. Incorrect assessments of consumer interest in a concept can result in misallocation of development and production resources. Placement of ineffective marketing creative represents the expenditure of budget without ROI, but also risks detrimental equity effects for the brand or product being marketed.

We propose the Artificial Market methodology as a methodological technique to address screening needs of product and advertising research. The Artificial Market Methodology uses market inspired purchase opportunities to elicit perceptions of relative market strength within a prelaunch testing study or screening program. We believe that the Artificial Market, which incorporates both respondent receptivity to concept or stimulus and their perceptions of market receptivity, has the potential to crowdsource market wisdom into a more reliable measure of in market success.

There are potential advantages related to respondent focus and experience realized by the Artificial Market that strengthens its potential relative to other methodologies. Respondents participating in Artificial Market research programs are incented in a way that encourages engagement with the market and earnestness in response. The Artificial Market method has the potential to be informed by respondent information while also distributing group assessment information back to participants to inform their future considerations.

This research paper presents a demonstration of the technique. It will report about an exploratory execution of the technique with preliminary analytic comparisons to a commonly implemented alternative. The paper will then define a roadmap for future research and development.

2. Artificial Market Methodology

Broadly, the Artificial Market seeks to create a relative scaling of ideas or concepts within a bounded, interval scale. The resultant metric is a measure of perceived Market Potential for the concepts or ideas being tested. Participants engage in a series of purchase tasks that mimic market opportunities that participants in a traditional predictive market[1] or an adapted commodity market might make[2].

Artificial Market participants are not risking their own assets in the exercise, but they have the potential to benefit from the decisions they make within the purchase tasks. The exercise is repeated across participants who are exposed to a portfolio or flow of concepts at various market potential "offers" similar in function to a market-based price. Within basic estimation, purchase decisions across these opportunities are aggregated by concept tests, followed by an estimation of a Market Potential score which seeks to find the level at which the inflection point between the number of those willing to sell the concept equals the number wanting to buy is found.

2.1 The Basic Process

The process starts with a list of concepts or ideas to be tested. This can be composed of a static list of concepts in a one-time study or a continual flow of new stimuli contributing to a database from which items are drawn. Ideas are presented to respondents as brief concepts that are commoditized with randomly created "offer" points that represents the market potential for the tested item. This "offer" point is bounded by the scale of 5 to 95 for an initial presentation of concepts.

Artificial Market participants react to this decision moment by evaluating the concept and the "offer" point relative to how they perceive the aggregate market is likely to evaluate the concept. If the respondent feels the "offer" point is low compared to how they expect the market will evaluate the concept, then they will respond by indicating "Purchase if I Could." If they feel the concept is offered at an index higher than the market is likely to evaluate it, then they respond by indicating "Sell if I Could." Formally, using definitions from above, upon exposure to concept j, each participant i will seek to maximize his virtual portfolio by deciding:

- If $PW_{ij} > P$ then participant will "buy" the concept, response = 1
- If $PW_{ij} < P$ then participant will "sell" the concept, response = 0

For a participant these purchase opportunities are presented n at a time, across multiple screens, so a participant will see a number of concepts from the pool of available concepts at a variety of price points. As this process is repeated across multiple participants and subsequent "plays" per individual concept, choice data is collected for the portfolio of concepts across a range of "offer" points and purchase choices are stored within a database.

With sufficient data in hand, an initial estimation can commence. Data is aggregated by concept and coded such that a "Purchase if I Could" response is coded as a "1" and "Sell if I Could" is coded as a "0." This dichotomous response variable is the dependent variable in a limited dependent regression model where the "offer" price is the independent variable. Throughout our research, we have chosen to utilize logistic regression for this purpose.

The purpose of the logistic model is to find the equilibrium point at which the market as a whole is indifferent between the purchase and sell options. The logistic regression model easily lends itself to this goal. When the market is indifferent to the concept, 50% of the market is willing to purchase the concept and 50% is willing to sell the concept. The log odds ratio, (P/(1-P)), is zero leaving the model as the expression 0 = $b_o + b_1P^*$. Collecting terms, we find $P^* = -b_o/b_1$. This is the model's estimate of the Market Potential for the concept.

Once the market potential score is estimated for a concept, the decisions made by individual participants are evaluated for correct insight relative to the market of participants. If a respondent responded "Purchase if I Could" to a concept where the market potential score was higher than the offer, the difference between the score and the offer is computed and the difference is deposited as points in the participant's "bank" account. Likewise, if a participant responded "Sell if I Could" to a concept where the market potential score was lower than the "offer" then the difference is calculated and deposited as points in the participant's account.

As points accrue in a participant's account the points can be used to score a contest rewarding the participants making the best decisions or as a debit account for a rewards program for all participants. This incentive serves two purposes, it creates engagement with the process and encourages good faith efforts in the evaluations provided.

For the user of the artificial market, the calculated market potential index is used to scale the portfolio or stream of concepts on a scale that can be used to inform the research needs of stakeholders to the testing program. The Market potential index can be used to benchmark ideas to a set portfolio, historical norms or market benchmarks. Stakeholders can then use this information to advance concepts or stimulus to the next phase in product development or media placement.

2.2 A Learning Model

An enhancement to the simple Artificial Market is to install a learning step in the algorithm. After an adequate number of exposures per concept is obtained, less than would be needed for full estimation, a preliminary estimate of the concept potential score is obtained using a linear probability regression model. Where $P^* = (.5 - b_o)/b_1$. Additional offers for a concept are now made with an "offer" point for market potential narrowed around the preliminary estimate. The tested range for the "offer" point narrows as more exposures are acquired. This narrowing of range concentrates additional exposures to the immediate portion of the response curve where the indifference inflection point would be expected to be found. This learning portion of the algorithm is expected to reduce the number of overall exposures needed to precisely estimate the final market potential scores.

2.3 Learning by Participants

In the base model, participants are not aware of how the market is responding to the tested concepts until after the market potential scores are estimated and payouts to their credit accounts are scored and made. An adaptation of the Delphi method[3] can be made to distribute information amongst participants that encourages the participants to adjust decisions to the market. After preliminary market potential scores are computed for each concept, participants may have access to a "scoreboard" that reports the current estimate for market potential scores for concepts. Consistent with expectations of the Delphi method itself, participants can evaluate prior bid decisions to inform decisions for future purchase opportunities.

3. Testing the Artificial Market

Testing of the Artificial Market began with a beta test in the winter of 2020, adapted and expanded into the test efforts that are described below. The test study contained two cells of research. The Artificial Market and as a contrast point, another commonly utilized methodology, Maximum Difference (Best-Worst) scaling[4].

3.1 Stimulus Used in Testing

The test stimulus was a batch of concepts with the intent of identifying the best concepts of the batch. The stimuli come from a set of campus related product concepts created within an entrepreneurial class of sophomores at the home institution of the authors. Sixteen of these concepts were selected for inclusion. Each concept was described in a short, single paragraph of text that was prepared in a similar format across concepts.

3.2 Participants Recruited and Incentives

The audience for the test were students within the business school of the author's institution. Within this sample, we expected to find general knowledge of the concept creation process involved with the entrepreneurial class. Depending on class standing, students had either completed the class or were currently enrolled in the class. The research team reached out to professors within the school who then passed along information and a survey link about the study to their students who could opt in while taking the screening survey.

Participants understood that they were to be part of a research program that involved multiple phases of research and they were to be incented by relative performance in the Artificial Market portion of the study. Incentives were gift cards (of participant's choice) for the top three performers. Participants were invited to participate in all phases of the research.

3.3 Research Phase

Cell 1—MaxDiff

Cell 1 consisted of a single survey featuring a Maximum Difference or Best-Worst Scaling presentation of the tested stimulus. MaxDiff was chosen as the first cell of research in order to expose participants to all concepts included in the testing portfolio so that respondents could create internal expectations for the range of concepts they would experience in later research phases.

Participants were exposed to 12 choice tasks. Each task presented 4 concepts on separate survey screens, with the 16 concepts randomized across the tasks. Each concept was seen 3 times. Within a task, participants were asked which concept they were "Most Likely" to purchase and which they were "Least Likely" to purchase. Using this choice data, Bayesian estimation of a multinomial logit created the basis for evaluation.

Cell 2—The Artificial Market

Cell 2 Consisted of 4 invitations to participate in the Artificial Market. Each participation featured 5 purchase opportunities where each opportunity featured a randomly drawn concept and an offer price. Each respondent evaluated the concept using the following copy:

> *Please review the product above.*
> *When thinking about the "final product potential" score this product might obtain, if this product were offered at a score of <Insert test score> out of 100.*
> *If you think the "Final Market Potential" score should be higher than currently listed, then "Purchase it." On the other hand, if you think the final Market Potential Score should be lower than currently listed, then "Sell it." If you are correct, you will collect the difference between the score above and the score obtained in the market.*
> *Would you purchase or sell?*

Invitations to the market were spaced 3 to 4 days apart spread over a 2-week period. When data collection was complete, the data was prepared by concept with the market potential being estimated for each individual concept using logistic regression.

4. Results of Initial Testing

4.1 Data Collection

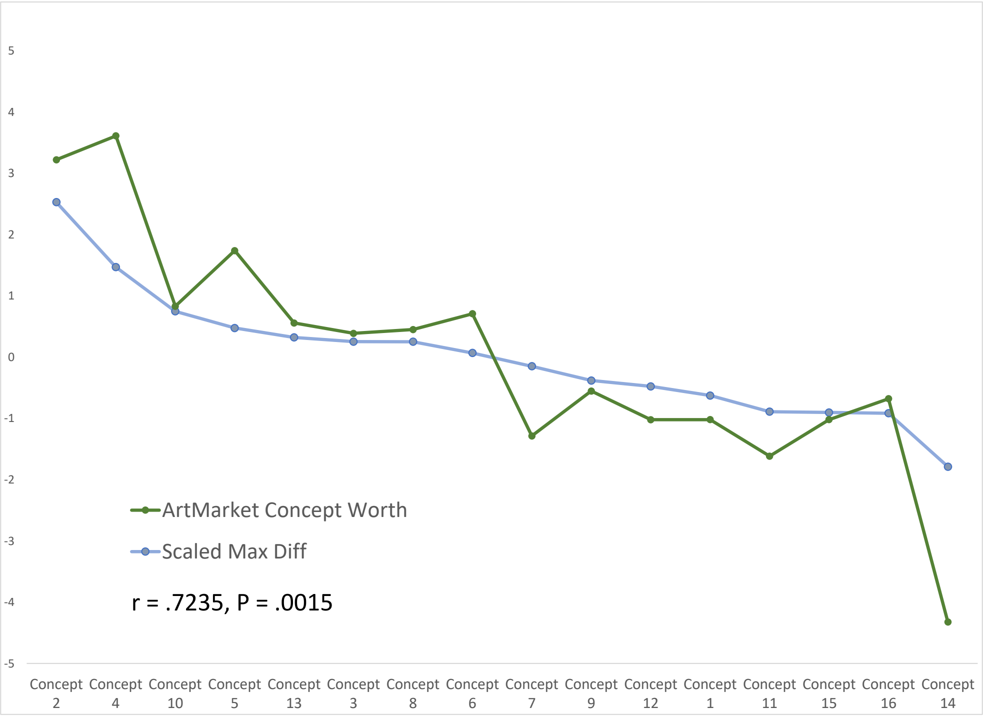
Data for exploratory testing was collected over a 7-week period from February 8 until March 26, 2021. 54 participants completed at least one round of Artificial Market tasks. 31 completed a single round of evaluations, 14 completed two and 9 completed three or more. 97 respondents completed a shadow cell of research using MaxDiff

(Best-Worst) scaling. Coefficients for both cells of research were estimated as described previously.

4.2 Comparison of Aggregate Results

Results of both methodologies are presented below in Exhibit 1. Model coefficients for both methodologies are normalized, paired by concept and sorted in rank order of the aggregate MaxDiff utilities. The correlation of the normalized scores between the two methodologies has a coefficient of .72 and is statically significant. There is general consistency in identifying winners and losers robust between the two methodologies.

Exhibit 1



4.3 Discussion

Overall, there is correlation between the two methodologies. However, there are multiple reversals or disagreements for the scoring of individual concepts. It is beyond the current data set to comment on these differences with confidence, but potential explanations are worth considering.

First, sample sizes for both cells were small. With these small samples the likelihood of sampling error in the estimates becomes a concern that cannot be excluded when considering differences in the resulting estimates. Second relates to the differences in reference that respondents are being asked to contemplate within the two methodologies as implemented in the current test.

Traditional methods of concept testing and screening, including MaxDiff scaling, expose individual consumers to a tested stimulus and then ask each respondent to evaluate the stimulus in a way that captures a measure of his or her expected utility. This process is repeated for a representative sample of potential consumers and then summarized into an aggregate score to provide insight for managerial decisions.

Within the Artificial Market, respondents are asked to evaluate concepts based on their perceptions of how the market will react to tested concepts. In taking such a speculative approach, the Artificial Market incorporates phenomena originally identified by Galton[5] now known as the "Wisdom of Crowds" to make predictions. Participants in the Artificial Market use prior witness of market wisdom to make inferences about new items being tested.

In summary, the MaxDiff scaling approach asks respondents to internalize an estimate of internal value for tested concepts. In contrast, the Artificial Market asks respondents to predict how other participants in the market will react. This difference of measurement dimension has the potential to create observed differences in the two tested cells.

5. Future Testing and Development

Several research questions remain to flesh out the justification of Artificial Markets as a testing methodology. These questions include:

1. How do evaluations from Artificial Markets compare relative to other tested methodologies as more robust samples are obtained?
2. In comparing an inward, utility maximizing evaluation to a wisdom of crowds approach to concept evaluation, which is the better predictor of market outcomes?
3. Can consistency between the Artificial Market and other methodologies be influenced by alterations in the evaluation basis for existing methodologies? Specifically, can the MaxDiff exercise be reworded to create an experiment that better lines up with the Artificial Market estimates?
4. How does a "learning" participant, one presented preliminary results via a Delphi method, affect future responses and aggregate estimates relative to those from a blind evaluation?
5. Does the "learning" algorithm for targeting "offer" points create efficiency in number of exposures required for the estimation of market potential?
6. Do novice respondents learn as they complete evaluations? Are later predictions a participant makes more reliable than earlier ones? Based on this answer, should early evaluations be given less weight?

6. Practical Implementation

Should testing continue to offer promise, multiple strategies may exist for the researcher to implement an Artificial Market testing program based on situation and need. The two main approaches tie back to the source and flow of the stimulus to be tested. The third is a hybrid approach.

6.1 Static Panel

A testing program featuring a standing panel of participants is reasonable when the concepts to be tested are a flow of items from a broad client pool who have some knowledge about what others in the pool desire. Examples of well-suited stimulus might include advertising copy or creative, or new product concepts.

In this form of implementation, there is an active flow of concepts available for testing which are sampled and presented to a standing panel until the necessary number of evaluations are met for each concept. At that point, that concept is removed from the active pool and the Market Potential score is calculated.

Once the score for a concept is calculated, it may be benchmarked to previously tested concepts. Care should be taken to consider the comparison set of concepts for similarity of category and consistency of presentation. Previously tested concepts which entered the market as part of an active advertising campaign or product launch are of special interest and can help form expectations of in-market success providing similar campaigns or product launch strategies are employed. As with any normative approach, care needs to be taken to ensure that referenced norms remain relevant over time.

In the static panel approach, aggregate wisdom of market tendencies are accumulated as panel members evaluate new items and, over time, observe payouts for concepts evaluated. This tethering to panel wisdom creates the opportunity to do normative benchmarking to previous ideas tested.

A standing panel allows for an incentive scheme where participants accrue points over time. Participants can cash out points according to either a standard redemption rule or a catalog of redemption items. This is likely to keep experienced participants engaged over the long term. This approach also allows for the training of new panelists. New panelists can be presented concepts and the first several evaluations need not be included in the final estimation of concept scores. Repeated exercises allow new panelists to get a feel for how the market values ideas. It is also possible to exclude panelists who perform poorly in their Artificial Market choices.

6.2 Point in Time Testing

An alternative approach, one utilized within this study, features a set of items to be tested relative to each other in a single study. Concepts tested are a list of ideas being screened for further development or potential claims or benefits used to position a product in the market.

Participants are recruited with the expectation follow up exercises to participate in the market. Special care for training must be incorporated into this type of design.

Participants may benefit from the posting of all tested concepts before the first invitation to the market. Similarly, using a learning approach to publish intermediate market potential scores for participant review provides the opportunity to fine tune responses to future market opportunities. Both these design points will be investigated in future research.

If benchmarking is a consideration, then benchmark concepts must be selected and presented within the set of new items being tested. These items should represent a range of performance for previously tested or launched items and be presented blindly, in a format consistent with new items. Market potential scores for new concepts can be evaluated relative to the scores obtained by the included benchmarks.

6.3 Multiple Concepts Sets over Time

Some organizations create waves of concepts that are comparable in content but created as sets separated by time. It is not realistic to maintain a static panel in this situation as lack of engagement is likely to result in defection from the testing panel. In these cases, it makes sense to recruit new participants for each wave of tested concepts. Recruitment strategies must be consistent wave to wave. Benchmarks of previously tested items can be carried over wave to wave. A creative researcher may be able to create a calibration scheme to transform ideas tested across waves into a single comparison space.

7. Conclusion

Overall, the Artificial Market has the potential to be a viable concept or creative screening tool for marketing or advertising researchers. While more testing is needed, in a preliminary experiment the Artificial Market has shown similarity to an established testing methodology while creating room for assimilating additional information that incorporates crowd-based wisdom. The Artificial Market is easily scalable and the most basic formulation of a testing study can be implemented with survey software that is readily available in market. More advanced enhancements require the development of a dedicated platform and app. Participants enjoy the experience and are easy to compensate in a way that keeps engagement and data quality up.

John Pemberton

Alfred Johnson III

Emily Gasparro

Alyssa MacDonald

Lauren Piskorski

References

Cowgill, Bo; Zitzewitz, Eric. “*Corporate Prediction Markets: Evidence from Google, Ford, and Firm X**.” Review of Economic Studies (2015) 82, 1309–1341.

Dahan, Ely; Kim, Adlar J.; LO, Andrew W.; Poggio, Tomaso; Chan, Nicholas. *“Securities Trading of Concepts (STOC).”* Journal of Marketing Research (2011) Vol. XLVIII (June 2011), 497–517.

Louviere, Jordan J. ; Flynn, Terry N.; Marley, Anthony Alfred John. ***Best-Worst Scaling Approach: Theory, Methods and Application***; Cambridge University Press (2015).

Dalkey, Norman; Helmer, Olaf (1963)*. “An Experimental Application of the Delphi Method to the use of experts.” Management Science.* ***9*** *(3): 458–467.*

Galton, F. “*Vox Populi.” Nature* **75,** 450–451 (1907).

ACBC VS. PARTIAL PROFILE CBC: A MARKET-BASED COMPARISON

FISHER LIU
DIAGAID
SHUMIN WANG
YI YOU
VIVO MOBILE COMMUNICATION CO.,LTD
DAPENG CUI
DIAGAID

ABSTRACT

Some users of full-profile ACBC have concerns on the length of the questionnaire, especially when there are many attributes and levels. They tend to find the choice questions of Partial-Profile CBC (ppCBC) containing a subset of the full profile to be easier and to fit in better with their thinking logic. Both methods are expected to lead to good prediction results.

In this study, we compared the predictive performances of ACBC against that of partial-profile CBC (ppCBC) in the setting of smartphone choice. We applied the methods with two independent samples of comparable sizes and profiles and followed the recommended designs as well as analysis procedures to examine whether there were significant differences between the two methods.

Our results show that ACBC did take longer to complete than ppCBC on average. The utilities estimated from both ppCBC and ACBC were highly comparable, and the attribute levels' preference ranking from both methods were similar. We also observed that the utilities of ppCBC looked like shrinkage estimators (toward less differentiated attribute importances) of those of ACBC.

However, the predictive accuracy of ACBC is much higher than that of ppCBC with respect to out-of-sample validation and real market sales volume. A smartphone manufacturer adopting ACBC results in its product optimization strategies achieved significantly better market growth than expected and outgrew its benchmark competitor in the new product series.

BACKGROUND

One of the top three smartphone manufacturers in China came to us for a conjoint research project on how consumers make purchase decisions when facing numerous smartphone features. The research was expected to help the client make better product configurations for both short-term and long-term strategies. More importantly, the client wanted the conjoint research to help them keep the leading position in the fierce competition in the Chinese market.

The long answer time of ACBC was the major concern, although the client highly valued the extra benefits gained from ACBC. As a result, the client wished to adopt a quick, simple, yet accurate choice experiment. Full-profile CBC was regarded as an alternative, but the client still thought it was too complex for respondents to complete when there were many attributes.

Inspired by the rather simple choice tournament task layout in an ACBC exercise, in which some unchanged attributes are greyed out, the client wanted to know if it was possible to replace ACBC with partial profile CBC. The research conducted by Michael Patterson and Keith Chrzan (Patterson and Chrzan, 2003) showed that ppCBC seems to work well when there were many attributes.

There were some debates between the client and us, but neither side could give convincing evidence to settle on either approach. In order to reach a consistent understanding, the client proposed to run a parallel comparison between ACBC and ppCBC.

RESEARCH DESIGN

We picked one of the client's major product lines for this parallel run experiment. Both ACBC and ppCBC experiments were created by using the same attributes and levels. Moreover, sampling criteria for both studies were the same and the choice data were collected from offline consumers. The final sample sizes for ACBC and ppCBC were 351 and 225.

1. Attributes and Levels

There were 14 attributes in this research. These attributes covered the most important smartphone features, including brand, price, CPU, storage, screen, camera, battery, charging speed, and biometric. The attributes and levels used are in Table 1.

Table 1: Attributes and Levels

Brand (*)	**CPU** (*)	**Rear Camera No.**	**Battery**
Brand H	CPU1	2	3500 mAh
Brand X	CPU2	3	4000 mAh
Brand Y	CPU3		
Brand V	CPU4	**Rear Camera Pixels**	**Fast Charge Speed**
	CPU5	24M	90 mins
Screen Size	CPU6	32M	60 mins
5.80"	CPU7	48M	30 mins
6.40"			
6.65"	**RAM**	**Front Camera No.**	**Biometric**
6.85"	4G	1	Face
	6G	2	Rear fingerprint
Screen Design (*)	8G	3	Screen fingerprint
Design1			
Design2	**ROM**	**Front Camera Pixels**	**Price**
Design3	128G	24M	Summed prices range from ￥1199 to ￥4999
Design4	256G	32M	
		40M	

(*) For confidentiality reasons, the real brand names, CPU models, and screen designs are not disclosed here.

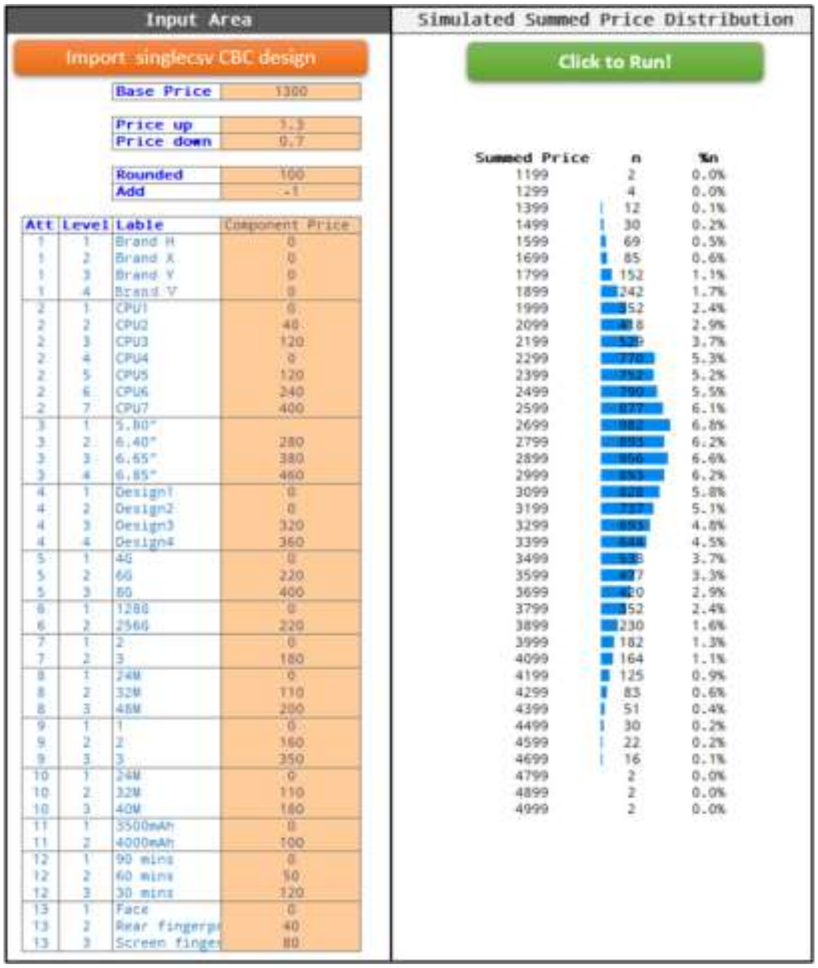
The price attribute in ACBC was designed as summed prices. The price of each concept was summed across the manifested levels of all attributes, and then it was varied with a random draw from anywhere from -30% to +30%. The summed prices were good at giving more reasonable prices than if just randomly drawn from some predefined price levels, even though there were always some difficulties for the client to give precise attribute level component prices.

To address the client's concern surrounding the summed prices, we created a summed price simulator to help the client flexibly adjust component prices and check the summed price distribution at the same time. To do this, a CBC design was created by using the same attributes and levels (not including price) and then the CBC design file was imported to an Excel simulator. The summed prices were calculated, in the Excel simulator, by applying the standard ACBC summed price formula (varying the summed prices +/-30%).

In this way, our client could modify the base price, component prices, and the total price variation range by themselves. And the Excel simulator would automatically simulate the summed prices for each product concept and display the simulated summed price distribution for checking.

This simulation ensured the final summed price distribution was consistent with the real market price distribution even if our guesses for some component prices were not fully accurate. The Excel simulator example is shown in Figure 1.

Figure 1: Summed Price Simulator


Input Area
Import singlecsv CBC design
Simulated Summed Price Distribution
Click to Run!

Base Price	1300
Price up	1.3
Price down	0.7
Rounded	100
Add	-1

Att	Level	Lable	Component Price
1	1	Brand H	0
1	2	Brand X	0
1	3	Brand Y	0
1	4	Brand V	0
2	1	CPU1	0
2	2	CPU2	40
2	3	CPU3	120
2	4	CPU4	0
2	5	CPU5	120
2	6	CPU6	240
2	7	CPU7	400
3	1	5.80"	
3	2	6.40"	280
3	3	6.65"	380
3	4	6.85"	460
4	1	Design1	0
4	2	Design2	0
4	3	Design3	320
4	4	Design4	360
5	1	4G	0
5	2	6G	220
5	3	8G	400
6	1	128G	0
6	2	256G	220
7	1	2	0
7	2	3	180
8	1	24M	0
8	2	32M	110
8	3	48M	200
9	1	1	0
9	2	2	160
9	3	3	350
10	1	24M	0
10	2	32M	110
10	3	40M	180
11	1	3500mAh	0
11	2	4000mAh	100
12	1	90 mins	0
12	2	60 mins	50
12	3	30 mins	120
13	1	Face	0
13	2	Rear fingerp	40
13	3	Screen finge	80

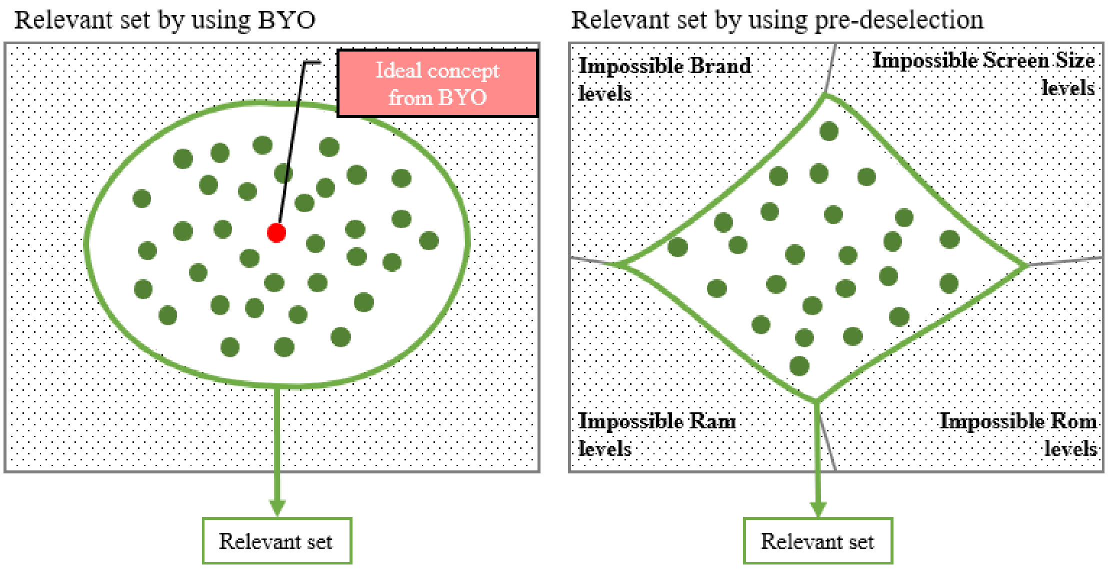
Summed Price	n	%n
1199	2	0.0%
1299	4	0.0%
1399	12	0.1%
1499	30	0.2%
1599	69	0.5%
1699	85	0.6%
1799	152	1.1%
1899	242	1.7%
1999	352	2.4%
2099	418	2.9%
2199	529	3.7%
2299	770	5.3%
2399	752	5.2%
2499	790	5.5%
2599	877	6.1%
2699	982	6.8%
2799	893	6.2%
2899	956	6.6%
2999	893	6.2%
3099	828	5.8%
3199	737	5.1%
3299	693	4.8%
3399	644	4.5%
3499	538	3.7%
3599	477	3.3%
3699	420	2.9%
3799	352	2.4%
3899	230	1.6%
3999	182	1.3%
4099	164	1.1%
4199	125	0.9%
4299	83	0.6%
4399	51	0.4%
4499	30	0.2%
4599	22	0.2%
4699	16	0.1%
4799	2	0.0%
4899	2	0.0%
4999	2	0.0%

2. ACBC Design

In our ACBC exercise, we did not use the standard Build-Your-Own (BYO) question (an option in the software allows you to drop this section). Instead, we asked select-type questions to let the respondent deselect/remove some key attribute levels from consideration, so these levels were not carried forward into the ACBC screening and choice tournament stages.

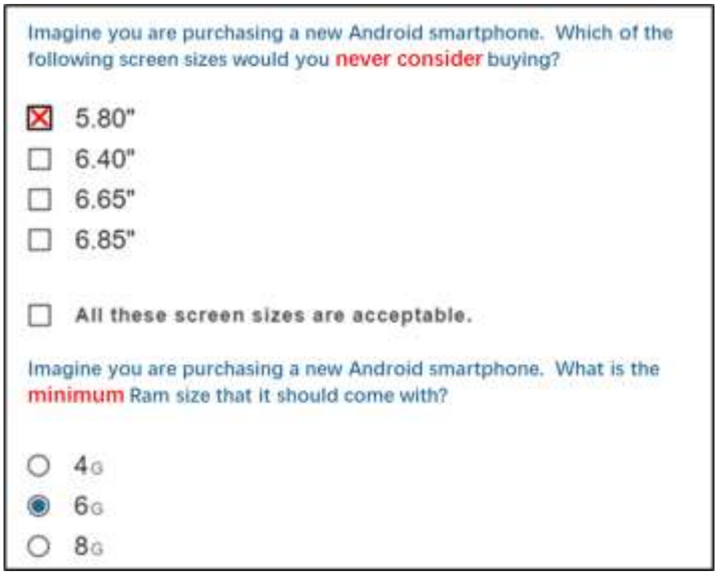
There were two reasons that we removed the BYO question in this study. First, some component prices were confidential, and we did not want to leak this information to respondents in the BYO question. Secondly, the key feature deselection process should be able to help the program home in on the acceptable level ranges even sooner for the key attributes. The differences between the standard BYO and pre-deselection on building relevant product concept set is displayed in Figure 2.

Figure 2: Relevant Set Defined by Standard BYO and Pre-Deselection Question



To avoid removing too many attributes' levels, we allowed respondents to remove levels only from Brand, Screen Size, Ram, and Rom attributes. The deselection question example is in Figure 3. From the client's understanding of the smartphone market, these four attributes were key buying factors for smartphone purchase decisions, and non-compensatory decisions were usually related to these four attributes. After the deselection stage, the desired key attributes' levels were passed into ACBC attribute level lists via Lighthouse Studio's constructed list function. For example, if a respondent deselected 5.80" from the screen size list, then only 6.40", 6.65" and 6.80" would be shown in subsequent ACBC questions. And if this respondent also selected 6G Ram as the minimal requirement, then 4G Ram would not be shown in the subsequent ACBC questions.

Figure 3: Deselection Question Example



Imagine you are purchasing a new Android smartphone. Which of the following screen sizes would you never consider buying?

☒ 5.80"
□ 6.40"
□ 6.65"
□ 6.85"

□ All these screen sizes are acceptable.

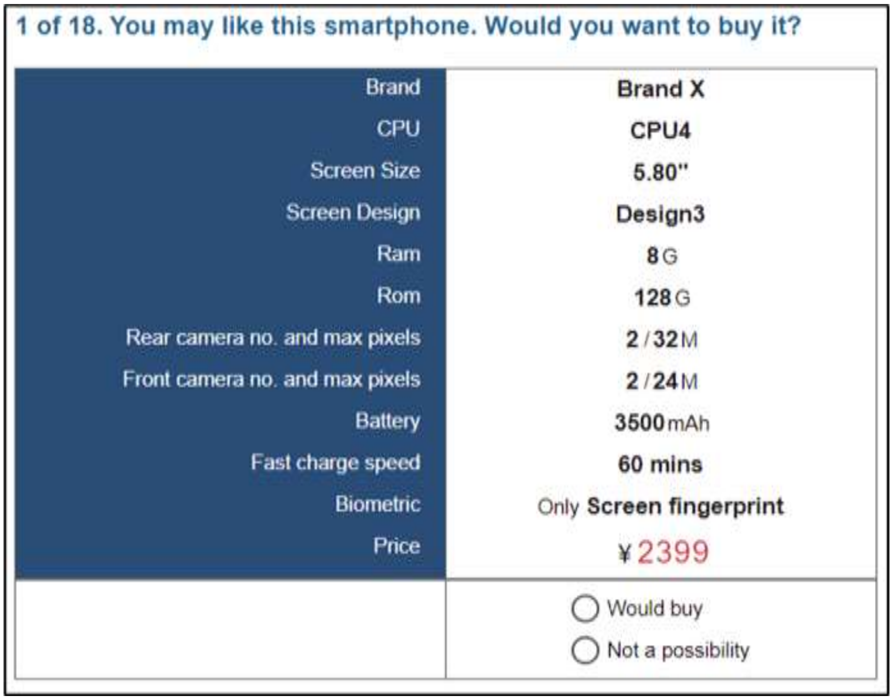
Imagine you are purchasing a new Android smartphone. What is the minimum Ram size that it should come with?

○ 4G
● 6G
○ 8G

In the ACBC screening stage, we displayed 18 screens of concepts and each screen just had ONE concept. The client hoped consumers would carefully evaluate each concept product for consideration, like what they normally would do in the real world, before making final decisions.

Given the fact it was an offline survey, we believed respondents could complete the 18 screening tasks smoothly. "Unacceptable" and "Must-have" question probes were also allowed in the screening stage, starting from the 10th screening task. A typical screening task in our study is in Figure 4.

Figure 4: Screening Task Example

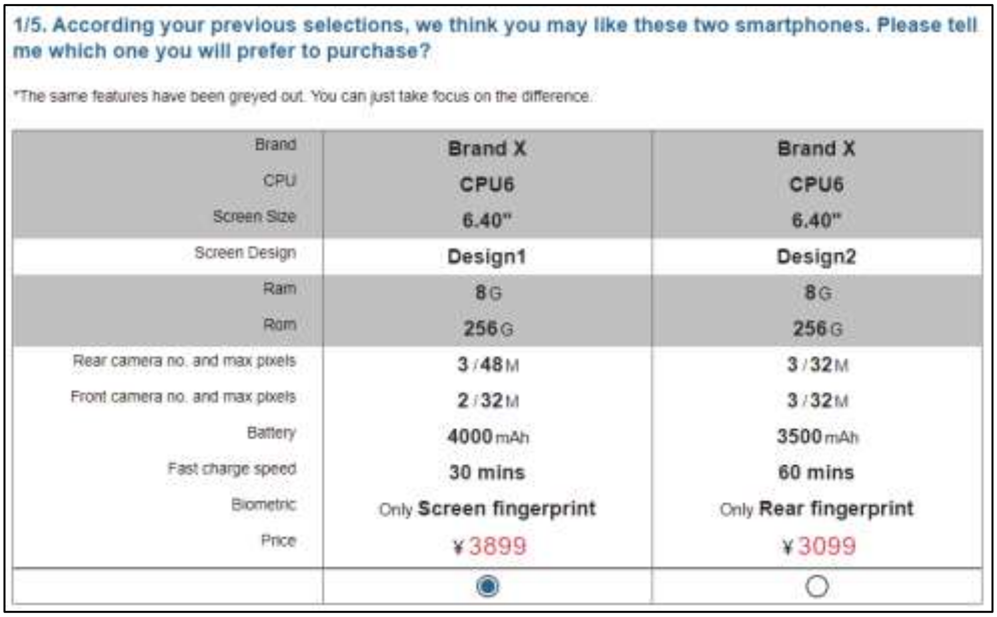


1 of 18. You may like this smartphone. Would you want to buy it?

Brand	Brand X
CPU	CPU4
Screen Size	5.80"
Screen Design	Design3
Ram	8G
Rom	128G
Rear camera no. and max pixels	2 / 32M
Front camera no. and max pixels	2 / 24M
Battery	3500mAh
Fast charge speed	60 mins
Biometric	Only Screen fingerprint
Price	¥2399
	○ Would buy ○ Not a possibility

In the ACBC choice tournament stage, our client preferred showing simpler choice tasks. A recent study by Martin Meissner, Harmen Oppewal, and Joel Huber suggested using pairs of concepts rather than ACBC's default triples when decisions are complex and difficult. We displayed just TWO concept products side by side and asked respondents make a choice in each task. There were 3-12 choice tasks in the choice tournament phase (to narrow down the one winning concept), depending on the number of screened-in concept products for each respondent. The choice task layout is in Figure 5.

Figure 5: Choice Task Example in ACBC Choice Tournament Stage



1/5. According your previous selections, we think you may like these two smartphones. Please tell me which one you will prefer to purchase?

*The same features have been greyed out. You can just take focus on the difference.

Brand	Brand X	Brand X
CPU	CPU6	CPU6
Screen Size	6.40"	6.40"
Screen Design	Design1	Design2
Ram	8G	8G
Rom	256G	256G
Rear camera no. and max pixels	3/48M	3/32M
Front camera no. and max pixels	2/32M	3/32M
Battery	4000mAh	3500mAh
Fast charge speed	30 mins	60 mins
Biometric	Only Screen fingerprint	Only Rear fingerprint
Price	¥3899	¥3099
	●	○

3. Partial Profile CBC Design

Generally, the partial profile CBC (ppCBC) design had the same attribute and level settings as in ACBC.

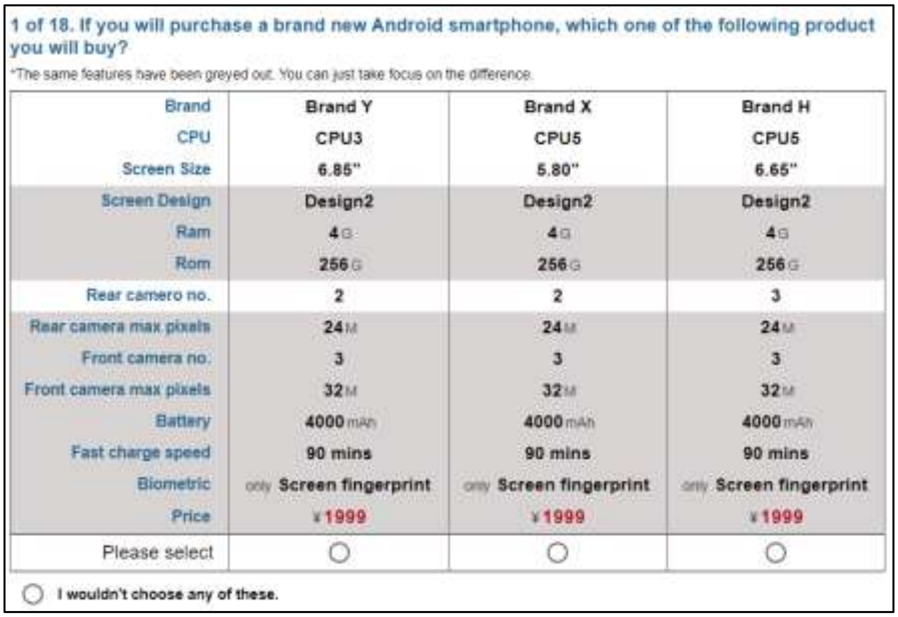
In the ppCBC design, 4 out of 14 attributes were varied in each task. The reason we used 4 active attributes in ppCBC came from the study done by Michael Patterson & Keith Chrzan (2003), in which they suggested using 3-5 active attributes in a typical ppCBC. In terms of the number of alternatives per task, we had thought of using paired comparison (two product concepts plus a "none" option), which means we would need to show quite a few tasks to achieve an acceptable design efficiency. Finally, we decided to let each respondent complete 18 tasks with 3 concepts plus a "none" option shown per task.

But we made some changes to the ppCBC exercise. We still showed "full-profile" concepts to respondents. Four attributes were formed from the real ppCBC design while the other attributes were kept at the same levels in each choice task. And we also generated a design for the greyed-out attributes to make sure their levels were varied in a balanced design across tasks.

We used the same summed price calculation for the price attribute to enable the comparability with the ACBC exercise.

We expected the modified ppCBC exercise would share some good characteristics with the full-profile approach (more complete context) while still being simple enough for respondents to complete. To some extent, it looked like the typical choice tournament task in ACBC, where tied attributes were "greyed out." A ppCBC choice task in our study is shown in Figure 6.

Figure 6: Partial Profile CBC Choice Task Example



1 of 18. If you will purchase a brand new Android smartphone, which one of the following product you will buy?

*The same features have been greyed out. You can just take focus on the difference.

Brand	Brand Y	Brand X	Brand H
CPU	CPU3	CPU5	CPU5
Screen Size	6.85"	5.80"	6.65"
Screen Design	Design2	Design2	Design2
Ram	4G	4G	4G
Rom	256G	256G	256G
Rear camero no.	2	2	3
Rear camera max pixels	24M	24M	24M
Front camera no.	3	3	3
Front camera max pixels	32M	32M	32M
Battery	4000mAh	4000mAh	4000mAh
Fast charge speed	90 mins	90 mins	90 mins
Biometric	only Screen fingerprint	only Screen fingerprint	only Screen fingerprint
Price	¥1999	¥1999	¥1999
Please select	○	○	○

○ I wouldn't choose any of these.

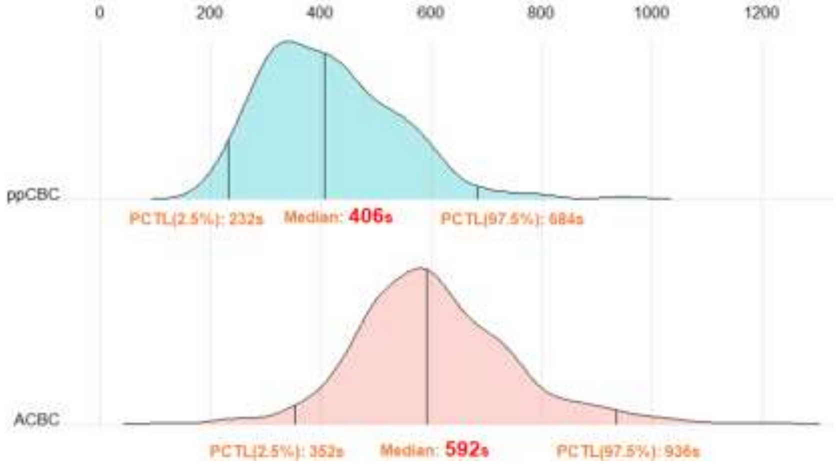
Result Comparison

1. Answer Time

On average, respondents took 6-7 minutes to complete the partial profile CBC and required about 10 minutes to complete the ACBC exercise. The answer time distribution is in Figure 7.

The actual mean answer time of ACBC was just 1.5 times longer than that of ppCBC, although based on previous research we had expected it would be at least 2-3 times longer than ppCBC.

Figure 7: Answer Time Distribution of ACBC and ppCBC

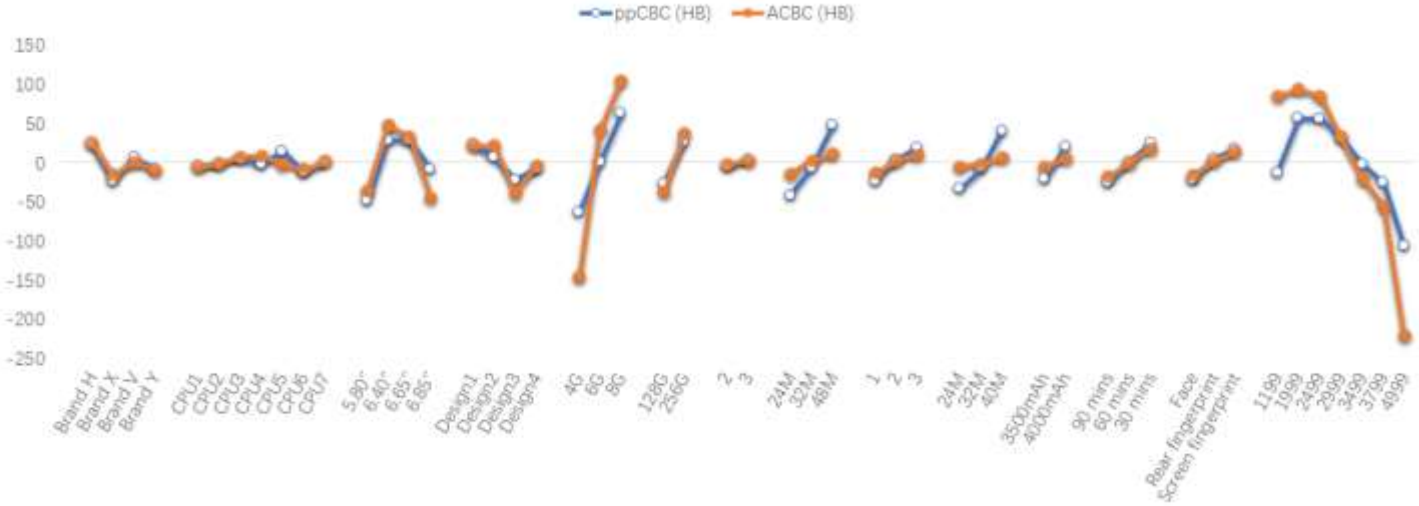


2. Utilities

We used an HB algorithm to estimate the attribute level utilities for both ACBC and partial profile CBC. Prices were piecewise coded when running HB, and seven key price point utilities along the piecewise function were reported for comparison. Both ACBC and ppCBC models used the full-profile coded design matrix.

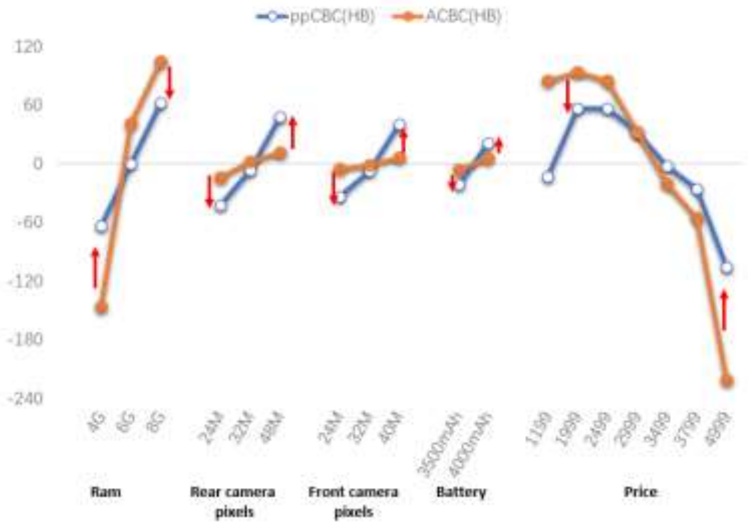
For each attribute, the utility ranking of its levels for ppCBC and ACBC were almost the same. The main-effect utilities per each attribute level are shown in Figure 8.

Figure 8: ACBC and ppCBC Main-Effect Utilities



However, we still observed some inconsistency between these two conjoint exercises. Ram and Price, the top 2 influential attributes in ACBC, became less impactful (relative to the other attributes) in ppCBC. Rear and Front camera pixels, and battery became much more important in ppCBC than in ACBC (again, relative to the other attributes). In general, ppCBC made the attributes' relative importances less extreme than ACBC. This overstatement and understatement of attribute impact is displayed in Figure 9.

Figure 9: Overstatement and Understatement of Impact



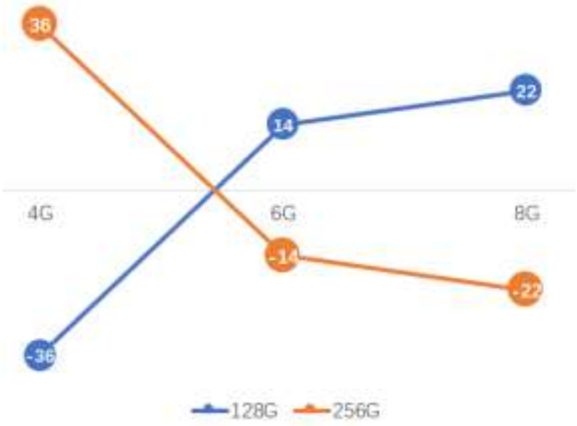
3. Ram and Rom Interaction

We also observed a significant interaction effect between Ram and Rom in ACBC (as modeled as additional parameters beyond the main-effect estimates). In ACBC, as the Ram size increased, the desirability for larger Rom (256G) went down. The interaction effect between Ram and Rom is shown in Figure 10.

It was our client that first asked us to pay attention to the possibility of Ram x Rom interaction effects. From our client's sales experience, they observed many consumers would rather purchase a smartphone equipped with larger Ram than with larger Rom. We tried to add this interaction effect into the model and found it was statistically significant in our ACBC choice data.

We also explored some additional potential interactions proposed by our client, but none of them were found to be as significant as the Ram x Rom interaction.

Figure 10: Ram & Rom Interaction Effect in ACBC



We tried to add a Ram x Rom interaction in the ppCBC model, but the interaction was found very weak in ppCBC. The ppCBC experimental design is less efficient than ACBC's design for detecting and modeling interaction effects. The interaction effects estimation of ACBC and ppCBC are in Table 2.

Table 2: Ram x Rom Interaction Effect in ACBC and ppCBC

	ppCBC (HB)	ACBC (HB)
4G x 128G	8	-36
4G x 256G	-8	36
6G x 128G	-5	14
6G x 256G	5	-14
8G x 128G	-3	22
8G x 256G	3	-22

Out-of-Sample Validation

Our client adopted the conjoint results for her own new product series configuration she intended to launch. Based on ACBC, ppCBC results and other research, they had decided on most of the new product features for the new product line launch, except the Ram & Rom configuration.

At that time, the client focused on scenarios with "8G Ram + 128G Rom" vs. "6G Ram + 256G Rom" competition. These two alternatives had similar prices but might have a cross attribute trade-off involving an interaction effect. Both alternatives were appealing to consumers.

We also tested the client's specific Ram-Rom tradeoff scenario within the ACBC and ppCBC choice simulators to obtain the predicted shares for these two alternatives. The ACBC prediction was based on the model that included the main-effect plus Ram x

Rom interaction effect while the ppCBC prediction was purely based on a main-effect model. These two models yielded quite different predictions.

The ppCBC simulation predicted a nearly 50-to-50 share prediction but the ACBC predicted the 8G + 128G alternative (71%) over the 6G + 256G (27%). The prediction result is in Table 3.

Table 3: "8G + 128G" vs. "6G + 256G" Simulation Result

Exercise	Sample Size	6G Ram + 256G Rom	8G Ram + 128G Rom	None
ppCBC	225	49%	46%	5%
ACBC	351	27%	71%	2%

These inconsistent predictions undoubtedly instigated some internal debates on the client side. Some people thought the ACBC prediction was right while others trusted the ppCBC prediction.

To justify which conjoint method was better for predicting the specific Ram-Rom tradeoff of interest, the client invested in a follow-up out-of-sample validation study. Three new groups of respondents were simply asked to make a choice isolating the tradeoff between the two attributes among "8G + 128G," "6G +256G," and "None" alternatives. The validation study was independently executed by a third party.

- Group1: online consumers (n=609)
- Group2: offline consumers (n=453)
- Group3: the client's offline-channel shop assistants (n=316)

The "8G + 128G" choice ratios for the 3 validation groups were 70%, 75%, and 76%, which were extremely close to the ACBC predictions (71%). The validation result is shown in Table 4.

Table 4: Out-of-Sample Choice Ratio Per Each Validation Group

Validation group	Sample Size	6G Ram + 256G Rom	8G Ram + 128G Rom	None
Online consumers	609	22%	70%	8%
Offline consumers	453	24%	75%	1%
Shop assistants	316	14%	76%	10%

This validation result gave our client more confidence to downsize or eliminate the 6G + 256G configuration in the new product line series.

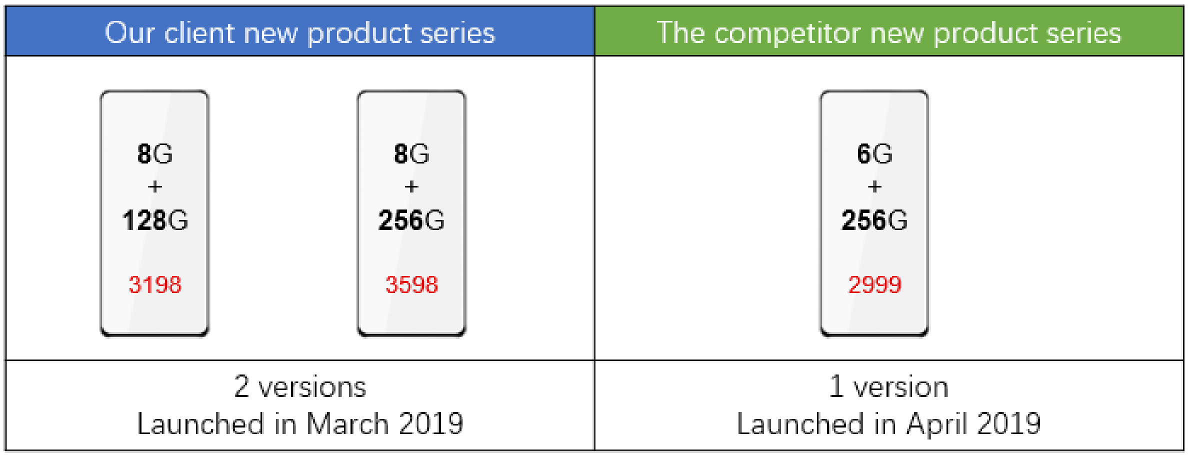
We think the following reasons helped ACBC make a more accurate prediction of the isolated Ram-Rom tradeoffs. First, ACBC is much more relevant and efficient than a ppCBC design. The adaptive mechanism, although a little bit tedious, really worked to hit the consideration space for most respondents. Secondly, ACBC's more efficient experimental design (with respect to interaction effects) was able to detect the unwillingness to upgrade Rom when Ram was good enough (the Ram x Rom interaction effect).

Real Market Competition

We also had an opportunity to verify the ACBC prediction results in real market competition. The conjoint studies we've described to this point were executed in November of 2018 and the client's new product series was launched 4 months later.

Our client's new product line had two versions: 8G Ram + 128G Rom and 8G Ram + 256G Rom. The major competitor's new product series just had one version: 6G Ram + 256G Rom. Both the client's new product series and the competitor's new product series were launched in adjacent months in 2019. See Figure 11 for the competition details.

Figure 11: Real Market Competition



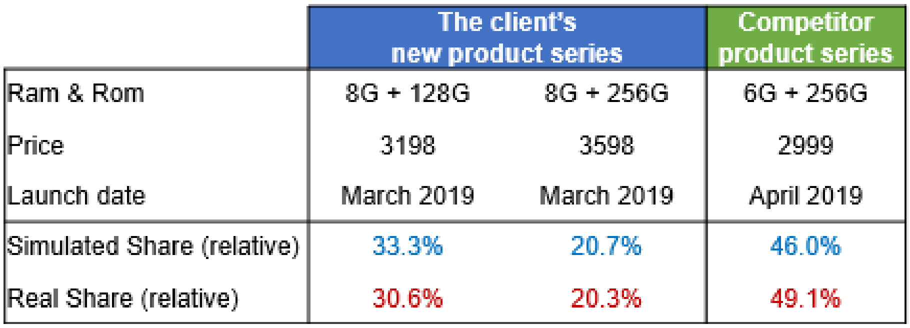
Our client new product series		The competitor new product series
8G + 128G 3198	8G + 256G 3598	6G + 256G 2999
2 versions Launched in March 2019		1 version Launched in April 2019

Both product series shared a lot of features except Ram, Rom, and price. These two brands had quite similar brand image and were positioned similarly in the market. Their offline channel coverage and sales capability were largely comparable, and their offline stores were almost side by side on the same street. Additionally, their production capacities were highly comparable. In short, the external effects for these two brands' competition can be somewhat ignored.

We also conducted a market simulation based on ACBC utilities. In this simulation, we applied the product availability factor (multi-store adjustment) to mimic the real market situation. Given the same production capacity, a multi-product strategy may result in more "out of stock" opportunity than single-product strategy, especially when the major sales of our client came from offline channels.

After checking the real historical sales data, our client confirmed ACBC's simulated relative preference share was remarkably close to the real market sales ratio. Our client praised ACBC highly for its market simulation predictive power. Although the client's new product was priced even higher than the competitor's product, they in the end achieved 16 percent more revenue than the competitor's product. The relative predicted shares from the ACBC simulator and the observed real market shares are in Figure 12.

Figure 12: Relative Simulated Share and Real Market Share



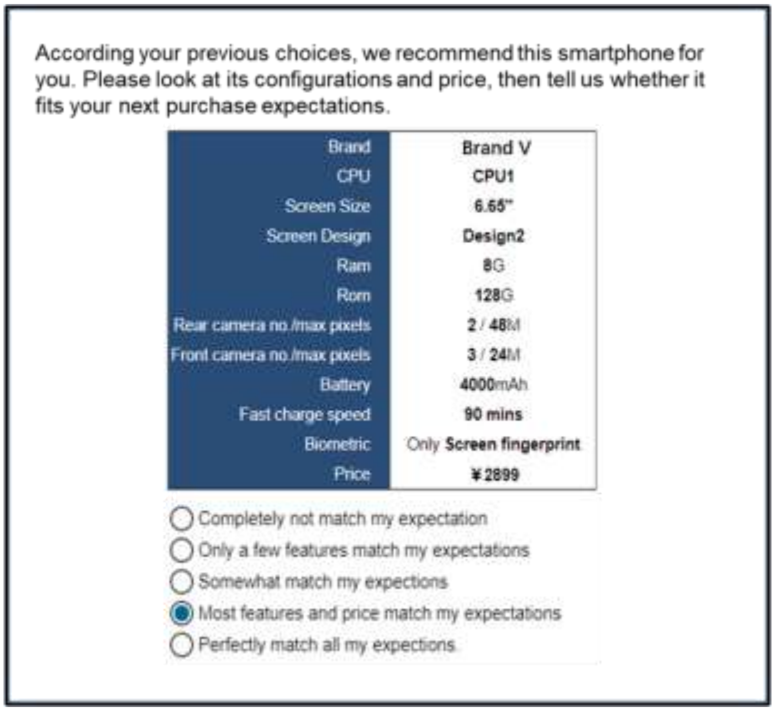
	The client's new product series		Competitor product series
Ram & Rom	8G + 128G	8G + 256G	6G + 256G
Price	3198	3598	2999
Launch date	March 2019	March 2019	April 2019
Simulated Share (relative)	33.3%	20.7%	46.0%
Real Share (relative)	30.6%	20.3%	49.1%

To conclude, the out-of-sample validation question (that isolated the Ram-Rom tradeoff) may only serve to verify a single aspect of the competition. In contrast, ACBC involving all 14 attributes empowered the client to incorporate the full picture of the features and competition and allowed us to answer many more "what-if" scenario questions.

ACBC Deep Dive

When we reviewed this study, we also did some exploration of the ACBC data. Actually, at the end of the ACBC questionnaire section, we took the ACBC winning concept and asked each respondent to evaluate the winning concept's fitness against their own expectations. This fitness assessment question example is in Figure 13.

Figure 13: Fitness Assessment for the ACBC Winning Concept



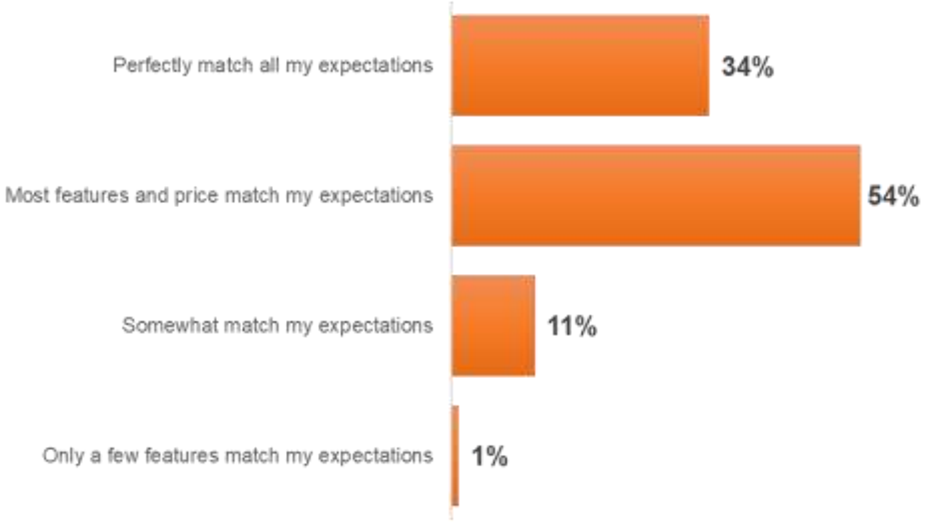
According your previous choices, we recommend this smartphone for you. Please look at its configurations and price, then tell us whether it fits your next purchase expectations.

Brand	Brand V
CPU	CPU1
Screen Size	6.65"
Screen Design	Design2
Ram	8G
Rom	128G
Rear camera no./max pixels	2 / 48M
Front camera no./max pixels	3 / 24M
Battery	4000mAh
Fast charge speed	90 mins
Biometric	Only Screen fingerprint
Price	¥2899

○ Completely not match my expectation
○ Only a few features match my expectations
○ Somewhat match my expections
● Most features and price match my expectations
○ Perfectly match all my expections

The claimed fitness was exceedingly high. Nearly 90% of respondents thought the ACBC winning product fit their expectations very well. See Figure 14.

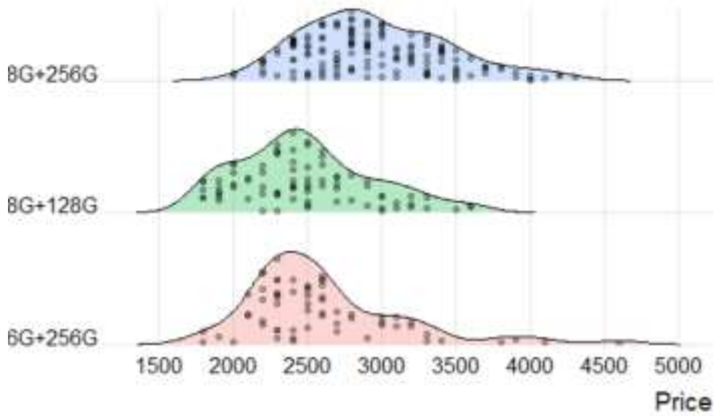
Figure 14: Claimed Fitness of ACBC Winning Concept

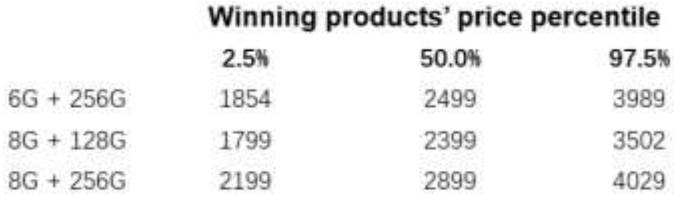


This result also showed that the dynamic ACBC process could help to identify the optimal or near optimal concepts at the individual level, even without modeling the choice data.

The winning concepts' median price by different Ram & Rom combination also make sense from the client's perspective. See Figure 15.

Figure 15: ACBC Winning Concept's Price Distribution



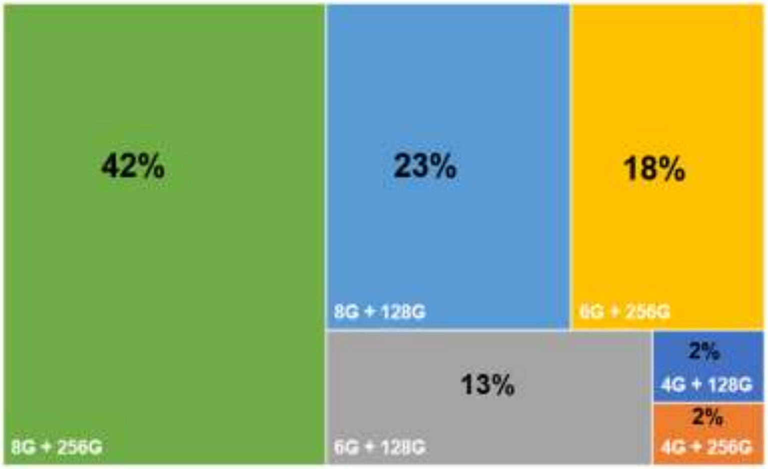


Winning products' price percentile

	2.5%	50.0%	97.5%
6G + 256G	1854	2499	3989
8G + 128G	1799	2399	3502
8G + 256G	2199	2899	4029

We also reviewed the Ram and Rom configuration count among all winning products. "8G + 256G" and "8G +128G" products were the top 2 winners, followed by "6G + 256G" product. See Figure 16.

Figure 16: Ram and Rom Count of ACBC Winning Concept



And we compared the Ram x Rom counts result from ACBC winning products with the Ram x Rom counts result from the pre-deselection questions. The result is shown in Table 5.

Table 5: Ram x Rom size of ACBC Winning Products by Self-Claimed Acceptable Ram x Rom

Pre-deselection	ACBC winning concepts					
kept in levels	4G+128G	4G+256G	6G+128G	6G+256G	8G+128G	8G+256G
4G+128G	18.2%	15.9%	15.9%	11.4%	**22.7%**	15.9%
4G+256G	17.8%	15.6%	15.6%	11.1%	**22.2%**	17.8%
6G+128G	3.7%	3.2%	20.1%	**26.0%**	**25.1%**	21.9%
6G+256G	3.4%	2.9%	18.5%	**26.5%**	**23.1%**	**25.6%**
8G+128G	2.9%	2.5%	15.8%	20.4%	**28.7%**	**29.7%**
8G+256G	2.3%	2.0%	12.5%	17.9%	**22.8%**	**42.5%**

This table also shows some inclination to upgrade Ram to 8G rather than to 6G. Those who accepted low Ram size (4G) or high Ram size (8G) would like to choose an 8G Ram product as the best one. Those who accepted medium Ram size (6G) would largely choose a 6G Ram as the best one but they also like 8G Ram products very much.

Conclusions

In this study, ppCBC was proved to be less accurate in prediction than ACBC when there were many (14) attributes. ppCBC overestimated the impact of some negligible attributes while underestimating some important attributes. An important interaction between Ram and Rom could not be detected either at the aggregate level or individual level in ppCBC.

A tailored key attributes level deselection process can replace the standard BYO under some circumstances and help the ACBC program quickly approach a more meaningful solution space for each respondent. But researchers should be cautious to include these key attributes in the design to avoid eliminating some meaningful attribute levels at the very beginning. The dynamic mechanism in ACBC and its resulting experimental design was helpful in identifying meaningful interactions between attributes.

ACBC worked much better than ppCBC in out-of-sample validation, as well as in market simulation. In a perfectly competitive market, where the external effects were held at the same level for different brands, as shown in this study, the ACBC model can accurately predict the real market share.

Considerations for Future Research

In this comparison study, the ppCBC just had 4 out of 14 attributes varied in the design. When we reviewed this study, we thought the ratio (4/14) was too low to get enough information to detect interactions. And the fewer the active attributes, the more likely respondents are to overrate some negligible factors.

Some previous studies (e.g., Keith Chrzan and Michael Patterson's presentation in the 2003 Sawtooth Software conference) have explored the optimal number of attributes shown in ppCBC (main effect model); still some experiments are needed to explore the optimal number of attributes shown in ppCBC when there are non-compensatory behaviors or interactions. We thought the constructed attribute list CBC, which let respondents first choose several most influential attributes and then just show these considered attributes in conjoint task, would also be worth trying to improve the efficiency and relevancy of the ppCBC design.

And we will also need to examine the held-constant attribute contribution in our ppCBC compared to the zeroing out the held-constant levels in the design matrix. In our study, we shaped the ppCBC choice task as a full-profile one with "none" option included. Without a "none" concept, the held constant (greyed out) levels across attributes do not contribute any information to utility estimation, although they provide more context for the respondent as the respondent answers the choice tasks. But, with a "none" concept in the choice task, the held-constant attributes indeed DO contribute some information to the design and for utility estimation. But we still need some empirical studies to know how much this method can improve the design efficiency or prediction power of ppCBC.

Fisher Liu

Shumin Wang

Yi You

Dapeng Cui

REFERENCES

Meissner, Martin, Harmen Oppewal, and Joel Huber (2016). How Many Options? Behavioral Responses to Two versus Five Alternatives per Choice. Sawtooth Software Conference Proceedings 2016.

Orme, Bryan K. and Keith Chrzan (2017). Becoming an Expert in Conjoint Analysis: Choice Modeling for Pros.

Patterson, Michael and Keith Chrzan (2003). Partial Profile Discrete Choice: What's the Optimal Number of Attributes. Sawtooth Software Conference Proceedings 2003.

Peitz, Megan, Mike Serpetti and Dan Yardley (2020). A researcher's guide to studying large attribute sets in choice-based conjoint. Sawtooth Software Conference Proceedings 2020.

Replication of Known Segment Structure and Membership: A Data-Driven Comparison of Robust Partitioning Methods for Metric Data

Keith Chrzan
Sawtooth Software, Inc.
Joseph White
InMoment

Background

Analysts frequently find themselves doing segmentation and over the years the Sawtooth Software Conference has seen a steady stream of presentations on segmentation methods. Despite, or perhaps because of, the plethora of segmentation methods available, analysts sometimes find themselves unsure about which methods to use. Herein we seek to reduce some of that uncertainty.

Some Things We Know

Under certain data conditions marketing scientists have a good idea how to choose segmentation methods. For example, when basing our segmentation on the results of a choice-based conjoint or MaxDiff experiment, latent class MNL is the way to go, as two papers in the *2019 Sawtooth Software Conference Proceedings* show convincingly (Eagle and Magidson 2019, Lyon 2019). The same holds true if we want to mix data from choice experiments with non-choice data: go with latent class MNL.

We also know that if we want to build segments from survey, behavioral or other data that mixes metric, binary, and unordered categorical data, we need a method that accommodates data of mixed scales. One such method, Partitioning Around Medoids (PAM) creates a common dissimilarities matrix for mixed variable types using the Gower statistic (Kaufman and Rousseeuw 1990, Retzer 2020). More commonly analysts use the Latent Gold software package, which allows latent class analysis of mixed variable types (Magidson and Vermunt 2002).

Some Things We Don't Know

For one very common situation, however, we don't know which of several segmentation methods works best: when we have some number of arguably metric survey (or other) measures like count, percentage, rating scale, and other ordered categorical variables. This may seem odd, because having a set of metric basis variables may seem like the archetypal case of segmentation. Very many methods could apply, ranging from simple k-means or PAM to any of the abundance of hierarchical clustering methods to more robust segmentation algorithms like model based (or latent class) clustering and methods that run many replications of analysis like convergent k-means, or that combine various segmentation solutions, like cluster ensembles analysis.

OBJECTIVE

To help analysts decide among the plethora of segmentation methods in this common use case, we conduct an experiment using artificial data with known segment structure: we know both the number of segments in our data and we know to which segment each record belongs.

RESEARCH PLAN

Data

We create artificial data sets in which respondents are members of non-overlapping segments each with a centroid located in a truncated multivariate normal distribution, as has been done in past studies (Milligan 1980, Milligan and Cooper 1985, Sawtooth Software 2013). We will independently vary five aspects of segment structure we think might affect the relative success of the different segmentation methods, for a total of 48 cells in our experiment:

- Number of segments: three versus five.
- Number of independent dimensions: segments may vary along three or five dimensions. Prior researchers noted that additional dimensions along which segments differ become redundant and make it too easy for segmentation methods to place respondents into the correct segments, so we will avoid that path.
- Alignment of basis variables with dimensions: here we have three conditions:
    - Each dimension is measured with a single variable
    - Each dimension is measured with five variables
    - Each dimension is measured with one variable while four "masking" variables uncorrelated with cluster membership are also included
- Amount of separation between segments: half of our datasets are separated by an average of one within-cluster standard deviation while the other by 0.5 times the within-cluster standard deviation.
- Relative segment sizes: half of our data sets feature equal segment sizes, but the other half limit the Kth segment to 1/K the size of segment 1.

To reduce the chances that one-off oddities influence our results we created 100 replicates of each of our 48 design cells. As described below, we then looked at the range of difficulty in those 100 replicates and selected 10 that spanned the range to use in our final analyses.

Note that what we're doing is in many ways an ideal situation for segmentation: our data sets have cluster structure built into them and the clusters differ meaningfully on several dimensions for which we have measures. Discovering these segments will automatically capture the meaningful structure that exists in the data sets. This may not be the case in empirical data sets where the segments that may be most useful for marketing purposes may not be the ones with the best fit in statistical terms.

Moreover, we have deliberately left the number of dimensions small, because additional dimensions that identify the segments in artificial data sets make the task too easy. Many empirical applications attempt to segment on many more dimensions, resulting in the kinds of problems discussed by Nowakowska and Retzer in this volume.

Segmentation Methods to Investigate

From experience and anecdotal evidence we believe that many simple segmentation methods produce unreliable results—they may be highly sensitive to starting points (or even to the order of cases in the data file) or they may tend to produce very different results from minor changes to software settings. We therefore focus our comparison on five robust partitioning methods:

- Convergent k-means cluster analysis: available in Sawtooth Software's CCEA package
- Cluster ensembles analysis: also available in CCEA or R
- PAM: available in R
- K-means with a search for the best solution from among 1,000 sets of starting points: available in R
- Model-based clustering (AKA finite mixture modeling or latent class analysis): available in Latent Gold and in R.

Once we started creating the data sets as described below it occurred to us that we could also test a couple of other closely related topics. For example, R has a package called NbClust that generates a consensus recommendation for the number of segments in a data file, based on the results of 30+ criteria suggested in the academic literature (Charrad et al. 2014). In addition, the Sawtooth Software segmentation methods report a reproducibility statistic that may be used to identify the correct number of segments, but computing a silhouette statistic for the Sawtooth Software solutions was an easy addition to our analysis, so we were able to test that as well. We tacked these additional comparisons to our analysis plan in the hope of adding more value to our project.

Other topics of possible interest (e.g., heteroskedasticity of basis variables, presence of outliers) are best handled during data inspection, cleaning, and transformation prior to cluster analysis, so we do not include them in our experimental treatments.

Success Criteria

We assess, across data conditions, each of the five segmentation methods.

- How well each method performs in terms of recovering the correct number of segments and
- For the correct number of segments, how each method fares in terms of placing respondents in their true segments.

We will also be able to note whether some of our segmentation methods work in some data conditions differentially better or worse than the other segmentation methods.

Data Generation

In order to assess the competing algorithms, we generate 100 synthetic data sets for each of the 48 cells in the experimental design, for a total of 4,800 unique data files. The data files are generated to ensure complete separation between clusters by selecting only records that are closest to their observed centroid in Euclidean distance. This is accomplished through an iterative process as follows.

1. Establish initial centroids from a multivariate normal distribution for each cluster in the design cell (3 or 5).
2. Generate 100,000 candidate records distributed multivariate normal around each of the initial centroids with design cell separation (0.5 or 1 standard deviation).
3. Select the number of records according to the design distribution (even or uneven).
4. Calculate the new shifted centroids from the resulting selection.
5. Remove any records that are no longer closest to their true shifted centroid and replace with unselected candidates from step 2.
6. Repeat step 4 and 5 until no replacements are needed.

Not all of our algorithms are especially conducive to automation, so we reduce these 100 replicates to a set of 10 for our full analysis. This we achieve by estimating solutions given the known number of segments using PAM, K-Means, and mclust (R), calculating the average rate of accurately classifying records into their true segments by replicate, and then selecting those associated with each decile of performance. This helps to ensure we present our methods with the full range of difficulty across the 100 initial replicates.

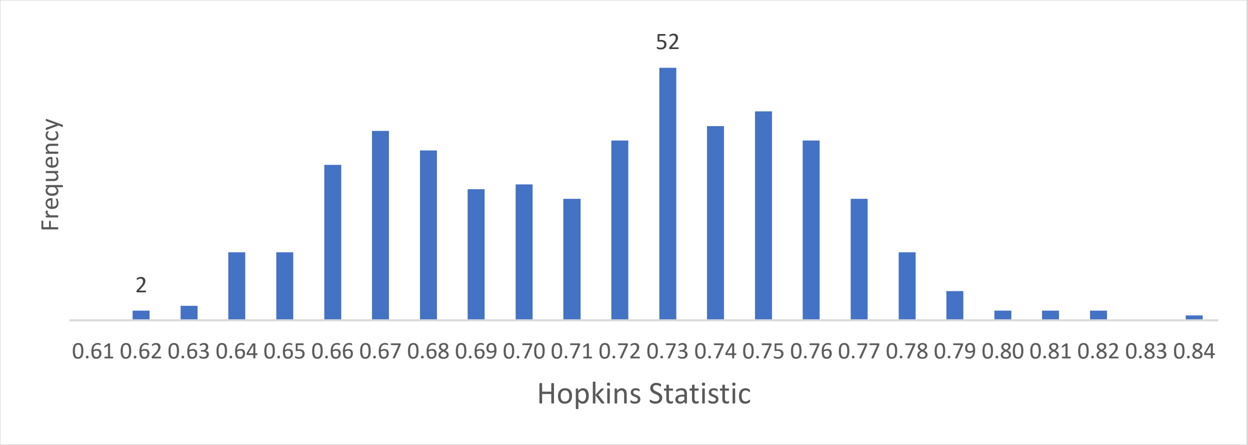
The process of ensuring complete separation in the data represents the best-case scenario for clustering. Across our replicates, we want to understand the range of difficulty presented to the clustering techniques. To do this, we look at the Hopkins statistic, silhouette plot, and VAT chart associated with each data file.

The Hopkins statistic is a measure of cluster tendency for a given data set that compares the true distribution of the data with a randomly generated uniformly distributed data set. The statistic compares average nearest neighbor distance in the real data to that for the uniformly random data, specifically:

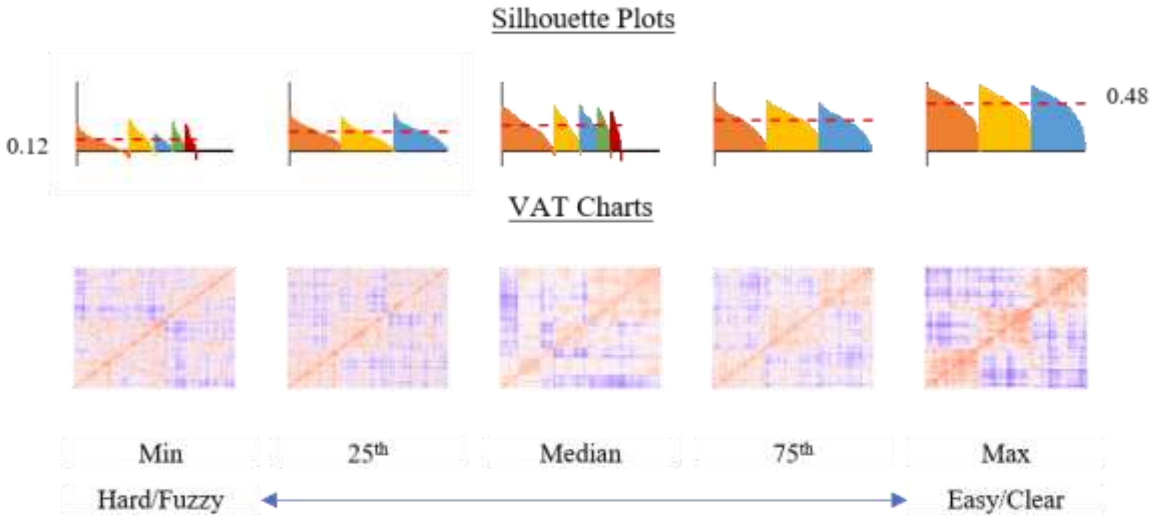
$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i},$$

where y_i is the distance between point i in the uniform random data and its nearest neighbor, and x_i is that for the actual data. If our actual data is uniformly distributed then the sum of x_i would be close to the sum of y_i, and H would be 0.5. As our data approaches perfect clusterability, so for each x_i there is an x_j such that $i <> j$ and $x_i = x_j$, the value of H will approach 1. The Hopkins statistic for our data sets range from 0.62 to 0.84, indicating some of the data sets are more conducive to clustering and some less.

In this way we cover a realistic range of difficulty in terms of the tendency of our data points to group together.



Inspection of the silhouette plots and VAT charts confirm the range of difficulty present in our synthetic data sets. Not surprisingly, the most difficult data set is associated with an uneven distribution of 5 clusters having lower separation (unit standard deviation), and the easiest is where there are 3 evenly distributed clusters with greater separation (0.5 standard deviation). The panel below shows the silhouette plots and VAT charts for the minimum, 25th percentile, median, 75th percentile, and maximum average silhouette width.



Negative silhouettes represent records that are more similar to neighboring than own cluster members, indicating fuzziness that should make cluster identification more difficult. The corresponding VAT chart further shows the lack of clear structure in this particular replication. Although one may intuitively think this contradicts the notion of complete separation it does not, as similarity with neighboring cluster members is distinctly different from distance to the neighbor centroid.

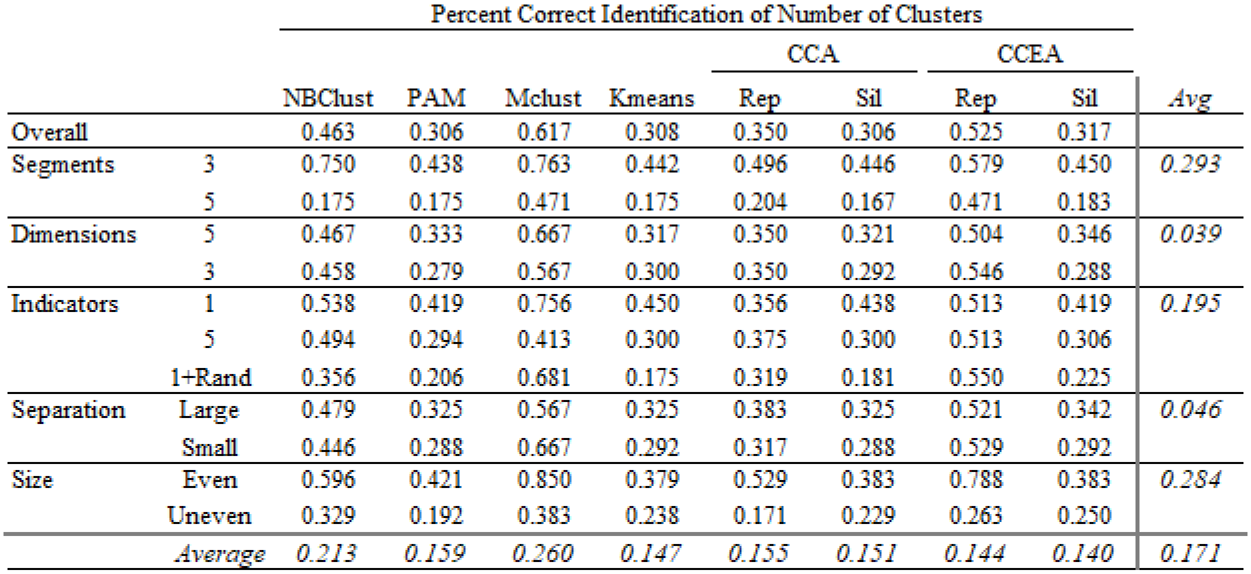
Correct Number of Clusters

The first assessment of our algorithms is with respect to their ability to correctly identify the number of clusters in the data. We leverage the recommended statistic for each of the techniques to identify the number of clusters that would result based on statistical analysis alone. The specific decision criteria are as follows.

- PAM and K-Means: run the algorithm for 2 to 8 cluster solutions and select the one with the greatest average silhouette width.
- mclust: model solutions for 1 to 8 clusters and identify the one with the best BIC.
- Sawtooth Software CCA and CCEA: generate solutions for 2 to 8 clusters and select the one with the greatest reproducibility statistic.

As much of the analysis is conducted in R, we take advantage of the NbClust package as another way to identify the number of clusters in the data. NbClust takes an ensemble approach, looking at 31 different statistics to reach a consensus estimate of the number of clusters present. Another addition for this phase of the analysis is to apply the maximum average silhouette rule to the CCA and CCEA solutions.

The table below shows the results for each technique employed to identify the number of clusters in the data.



Percent Correct Identification of Number of Clusters

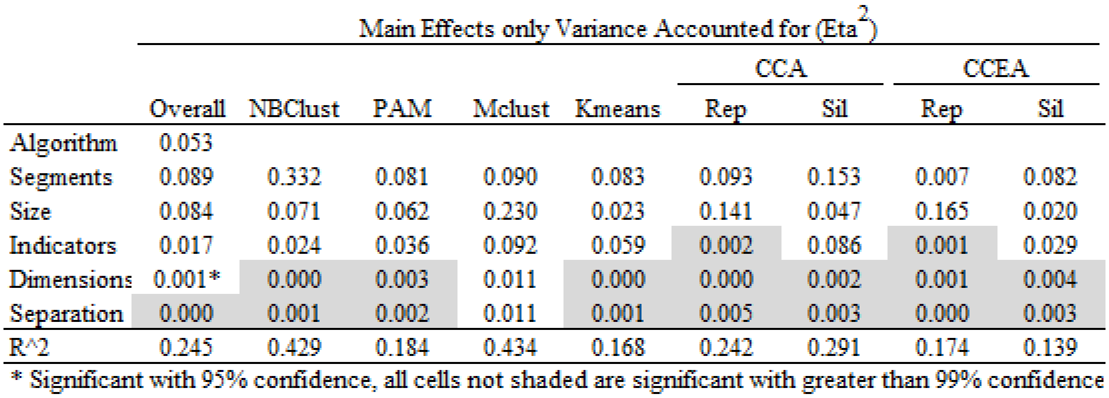
						CCA		CCEA		
		NBClust	PAM	Mclust	Kmeans	Rep	Sil	Rep	Sil	*Avg*
Overall		0.463	0.306	0.617	0.308	0.350	0.306	0.525	0.317	
Segments	3	0.750	0.438	0.763	0.442	0.496	0.446	0.579	0.450	*0.293*
	5	0.175	0.175	0.471	0.175	0.204	0.167	0.471	0.183	
Dimensions	5	0.467	0.333	0.667	0.317	0.350	0.321	0.504	0.346	*0.039*
	3	0.458	0.279	0.567	0.300	0.350	0.292	0.546	0.288	
Indicators	1	0.538	0.419	0.756	0.450	0.356	0.438	0.513	0.419	*0.195*
	5	0.494	0.294	0.413	0.300	0.375	0.300	0.513	0.306	
	1+Rand	0.356	0.206	0.681	0.175	0.319	0.181	0.550	0.225	
Separation	Large	0.479	0.325	0.567	0.325	0.383	0.325	0.521	0.342	*0.046*
	Small	0.446	0.288	0.667	0.292	0.317	0.288	0.529	0.292	
Size	Even	0.596	0.421	0.850	0.379	0.529	0.383	0.788	0.383	*0.284*
	Uneven	0.329	0.192	0.383	0.238	0.171	0.229	0.263	0.250	
	Average	*0.213*	*0.159*	*0.260*	*0.147*	*0.155*	*0.151*	*0.144*	*0.140*	*0.171*

The "Rep" and "Sil" columns for CCA and CCEA indicate results using either the reproducibility statistic or maximum silhouette rule for determining the number of clusters, respectively. The average column shows the mean difference between the best and worst effect level across techniques, and the average row indicates the mean effect impact by technique.

In general, the ability of our algorithms and statistics to identify the correct number of clusters is not good. Only mclust and Sawtooth Software's CCEA break the 50% mark. The maximum silhouette approach consistently performs poorly, only finding the right number of segments about 30% of the time, suggesting one might want to consider alternative approaches for K-Means and PAM. Sawtooth Software's reproducibility statistic outperforms silhouettes for both CCA and CCEA.

The number of segments, cluster size, and indicators have the greatest average impact on finding the right number of groups. NbClust in particular is hit hard by the number of segments in the data, and size is especially problematic for mclust, CCA, and CCEA. Interestingly, Sawtooth Software's CCEA using reproducibility is relatively invariant with respect to the number of indicators or the inclusion of masking variables. Across our experimental cells the techniques have similar average effects, with the exception of mclust and NbClust which appear to be more susceptible to these data challenges.

To further assess the impact of our effects on the ability to identify the correct number of clusters, we conduct an ANOVA with percent correct as the dependent variable. The results support the findings in the previous table that number of segments, cluster size, and the number of indicators have the greatest impact on success. The ANOVA results for the main effects only model are displayed in the following table of Eta^2 values.



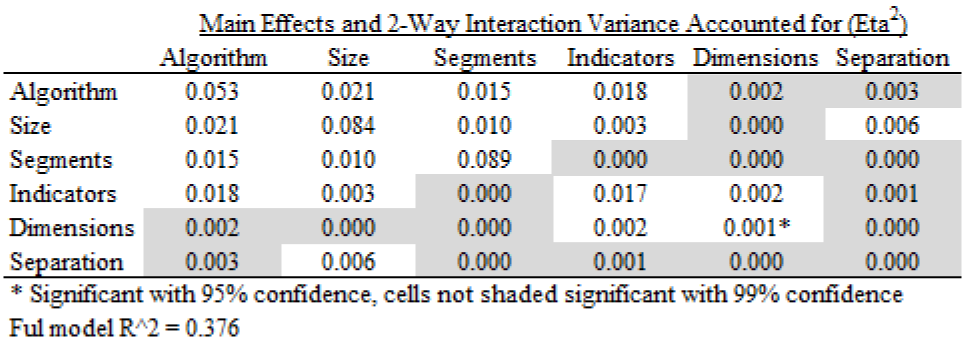
Main Effects only Variance Accounted for (Eta^2)

						CCA		CCEA	
	Overall	NBClust	PAM	Mclust	Kmeans	Rep	Sil	Rep	Sil
Algorithm	0.053								
Segments	0.089	0.332	0.081	0.090	0.083	0.093	0.153	0.007	0.082
Size	0.084	0.071	0.062	0.230	0.023	0.141	0.047	0.165	0.020
Indicators	0.017	0.024	0.036	0.092	0.059	0.002	0.086	0.001	0.029
Dimensions	0.001*	0.000	0.003	0.011	0.000	0.000	0.002	0.001	0.004
Separation	0.000	0.001	0.002	0.011	0.001	0.005	0.003	0.000	0.003
R^2	0.245	0.429	0.184	0.434	0.168	0.242	0.291	0.174	0.139

* Significant with 95% confidence, all cells not shaded are significant with greater than 99% confidence

The Eta^2 value represents the amount of explained variance associated with each effect. These are rescaled to reflect the decomposition of the R^2 across effects. The R^2 is thus the sum of the Eta^2 values. The number of segments in the data and cluster size are consistently the top contributors to the explained variance across techniques. The number of indicators is also a top contributor, except for Sawtooth Software's reproducibility statistic where the effect is not significant.

An all-interactions ANOVA is also run, with main and two-way interactions presented in the following table of Eta^2.



Main Effects and 2-Way Interaction Variance Accounted for (Eta^2)

	Algorithm	Size	Segments	Indicators	Dimensions	Separation
Algorithm	0.053	0.021	0.015	0.018	0.002	0.003
Size	0.021	0.084	0.010	0.003	0.000	0.006
Segments	0.015	0.010	0.089	0.000	0.000	0.000
Indicators	0.018	0.003	0.000	0.017	0.002	0.001
Dimensions	0.002	0.000	0.000	0.002	0.001*	0.000
Separation	0.003	0.006	0.000	0.001	0.000	0.000

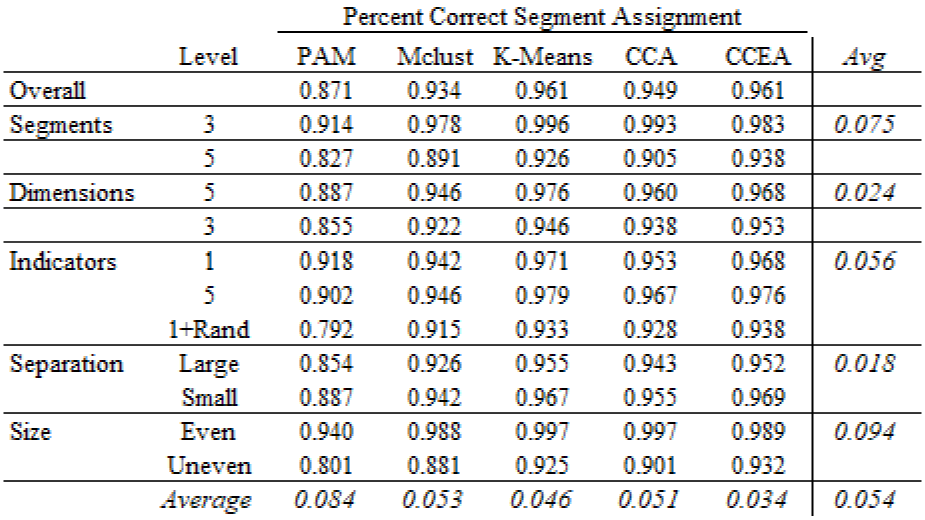
* Significant with 95% confidence, cells not shaded significant with 99% confidence

Ful model R^2 = 0.376

The full model improves our explanatory power by a little more than 50% with R^2 going from 0.245 to 0.376. The full model dominant effects are algorithm, cluster size, and number of segments; their main and two-way interactions accounting for 72.5% of the total variance explained. While we do find several significant higher order interactions, their contribution to the model is minimal, accounting for only 12% of the overall R^2.

Accuracy of Cluster Assignment

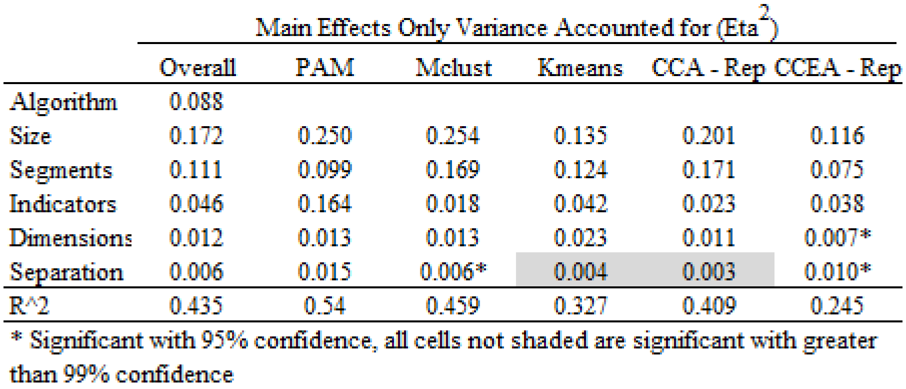
The second focus of our analysis is on how well each technique performs at accurately assigning respondents to their cluster if we specify the correct number of segments. It is expected that if the correct number of segments is known then our techniques should do well with classification. This, for the most part, is what we find as shown in the following summary table.



		Percent Correct Segment Assignment					
	Level	PAM	Mclust	K-Means	CCA	CCEA	*Avg*
Overall		0.871	0.934	0.961	0.949	0.961	
Segments	3	0.914	0.978	0.996	0.993	0.983	*0.075*
	5	0.827	0.891	0.926	0.905	0.938	
Dimensions	5	0.887	0.946	0.976	0.960	0.968	*0.024*
	3	0.855	0.922	0.946	0.938	0.953	
Indicators	1	0.918	0.942	0.971	0.953	0.968	*0.056*
	5	0.902	0.946	0.979	0.967	0.976	
	1+Rand	0.792	0.915	0.933	0.928	0.938	
Separation	Large	0.854	0.926	0.955	0.943	0.952	*0.018*
	Small	0.887	0.942	0.967	0.955	0.969	
Size	Even	0.940	0.988	0.997	0.997	0.989	*0.094*
	Uneven	0.801	0.881	0.925	0.901	0.932	
	Average	*0.084*	*0.053*	*0.046*	*0.051*	*0.034*	*0.054*

The robust K-Means and CCEA are clear winners in accurately assigning respondents to clusters with greater than 96% accuracy overall. However, each of our methods performs well overall, with the exception of PAM. This underperformance is due primarily to the inclusion of the masking variables and uneven segment size cells, which are especially detrimental to PAM relative to the other techniques. In general, we see a consistent story of number of segments, indicators, and segment size being the key challenges. Sawtooth Software’s CCEA is notably impacted the least by our design variables in terms of accuracy.

Similar to our analysis of identifying the correct number of clusters, we also conduct an ANOVA with accuracy as our dependent measure. The main effects only models are summarized below, again looking at Eta^2.



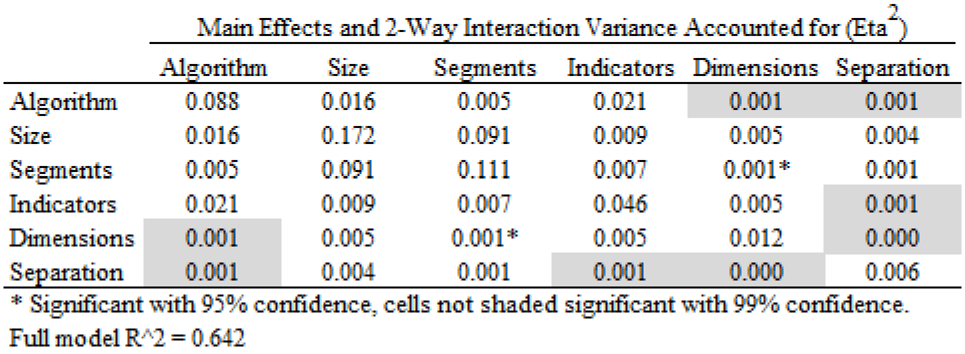
Main Effects Only Variance Accounted for (Eta^2)

	Overall	PAM	Mclust	Kmeans	CCA - Rep	CCEA - Rep
Algorithm	0.088					
Size	0.172	0.250	0.254	0.135	0.201	0.116
Segments	0.111	0.099	0.169	0.124	0.171	0.075
Indicators	0.046	0.164	0.018	0.042	0.023	0.038
Dimensions	0.012	0.013	0.013	0.023	0.011	0.007*
Separation	0.006	0.015	0.006*	0.004	0.003	0.010*
R^2	0.435	0.54	0.459	0.327	0.409	0.245

* Significant with 95% confidence, all cells not shaded are significant with greater than 99% confidence

Consistent with the performance summary, segment size has the greatest contribution to explaining differences in accuracy, followed in most cases by the number of segments. The notable outlier here is PAM, which as we saw in our performance results is hit particularly hard by the inclusion of irrelevant random variables in the base.

The all-interactions ANOVA again results in a roughly 50% improvement in the explained variance of the model overall, with the R^2 going from 0.435 to 0.642.



Main Effects and 2-Way Interaction Variance Accounted for (Eta^2)

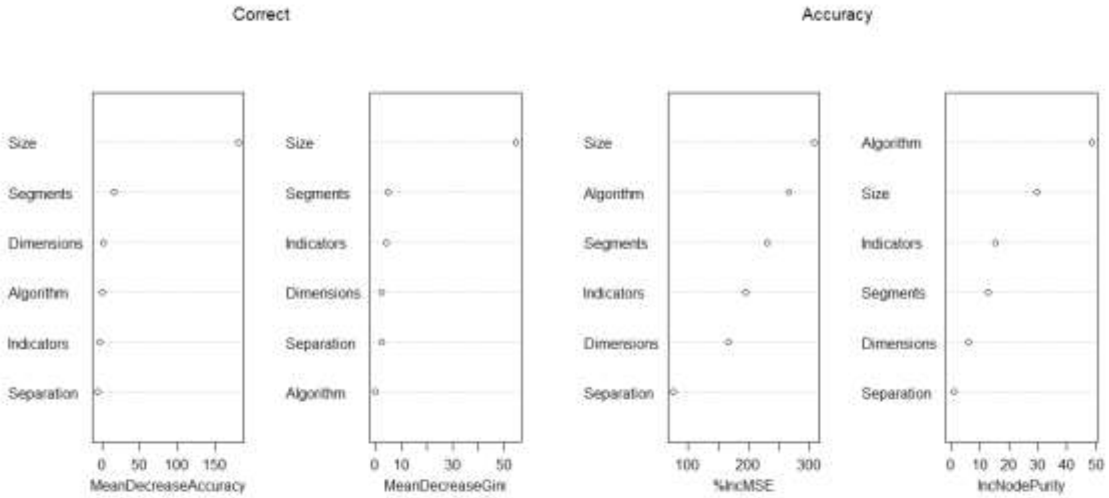
	Algorithm	Size	Segments	Indicators	Dimensions	Separation
Algorithm	0.088	0.016	0.005	0.021	0.001	0.001
Size	0.016	0.172	0.091	0.009	0.005	0.004
Segments	0.005	0.091	0.111	0.007	0.001*	0.001
Indicators	0.021	0.009	0.007	0.046	0.005	0.001
Dimensions	0.001	0.005	0.001*	0.005	0.012	0.000
Separation	0.001	0.004	0.001	0.001	0.000	0.006

* Significant with 95% confidence, cells not shaded significant with 99% confidence. Full model R^2 = 0.642

All but five two-way interactions are highly significant, and number of segments by dimensions is significant with 95% confidence. Several higher order interactions are also significant, but main and two-way effects account for a little more than 94% of the total explained variance. The biggest contributors, number of clusters and segment size along with their two-way interactions, account for 58.3% of the explanatory power of the model. Interestingly, choice of technique only contributes 14% as a main effect and just 20% including two-way interactions.

Relative Importance of Effects

The final analysis involves looking at how influential each of the design variables are in driving our ability to correctly identify the number of underlying clusters and accurately assigning respondents when the true number is known. We leverage the randomForest package in R to estimate models for both, each with 5,000 trees, default modeling parameters otherwise. The importance results are presented in the panel below.



Segment size balance is by far the largest driver of correctly identifying the number of segments, with algorithm selection playing a trivial role in success. Unfortunately, it would seem that unbalanced segments are likely to occur more often in practice, so there should be healthy skepticism about what our algorithms tell us.

However, if we do know the correct number of clusters, choosing the right algorithm becomes a top driver of accurately grouping respondents into their appropriate segment. And this is good news since it is the one aspect of our analysis that we can control as practitioners. Sawtooth Software's CCEA and the robust K-Means are clear choices for accurately assigning respondents once the number of clusters has been chosen.

Conclusions

None of the techniques we tested could reliably identify the correct number of segments present in a data set. Only mclust and Sawtooth Software's CCEA ensembles program managed even to get the number of segments right half the time. The design variables most harming the ability of our methods to find the right number of segments were (a) larger numbers of segments and (b) imbalances in segment size.

Fortunately, if we do manage to identify the right number of segments, all the algorithms besides PAM put more than 90% of respondents into the correct segments. The robust k-means (with 1,000 random starting points) available in R and Sawtooth Software's CCEA ensembles program performed best in terms of classifying respondents into the correct segments (assuming we select the correct number of segments).

For best results we recommend

- Use Sawtooth Software's CCEA (ensembles) or better yet mclust in R to identify how many segments are present in your data (and hope they point you to the right solution).
- Use R's k-means algorithm with 1,000 random sets of starting points or Sawtooth Software's CCEA (ensembles) to create segments.

Suggestions for Future Research

Future researchers might build on this research by focusing more on the variables we found to be important and eliminate some of the ones we found to be less important. Initially we thought we just scratched the surface in covering the topic of irrelevant "masking" variables, using only either zero or four of them and not doing any pre-analysis to discover them: it turns out that handling them by use of a procedure called cluster variable selection (Scrucca and Raftery 2018) works very well in our data sets, so this aspect of our study may not need to be repeated in future work.

We selected a handful of methods we consider robust and that we could readily access. Others might extend this analysis to other methods (as our discussant Andy Elder did with SPSS's 2-Step clustering procedure).

Keith Chrzan

Joseph White

References

Charrad, M., N. Ghazzali, V. Boiteau, A. Niknafs (2014) "NbClust: An R package for determining the relevant number of clusters in a data set," Journal of Statistical Software, 61(6): 1–36.

Eagle, T.C. and J. Magidson (2019) "Segmenting choice and non-choice data simultaneously: Part deux," Sawtooth Software Conference Proceedings, 247–280.

Fowlkes E.B and C. L. Mallows (1983) "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, 78:383, 553–569.

Kaufman, L. and P.J Rousseeuw (1990) *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

Lyon, D.W. (2019) "Comments on 'Segmenting choice and non-choice data simultaneously: Part deux,'" Sawtooth Software Conference Proceedings, 281–288.

Magidson, J. and J.K. Vermunt "Latent class models for clustering: A comparison with K-means," Canadian Journal of Marketing Research, 20: 37–44.

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,” paper presented at the 2021 Sawtooth Software Conference and included in this volume.

Retzer, J. (2020) “Painless & useful clustering of mixed mode data,” paper presented at the 2020 Sawtooth Software Conference.

Sawtooth Software, Inc. (2013) “CCEA V3,” downloaded from https://sawtoothsoftware.com/resources/software-downloads/convergent-cluster-ensemble-analysis.

Scrucca, L. and A.E. Raftery (2018) “clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R,” *Journal of Statistical Software*, **84**(1): 1–28.

BiCluster Identification & Profiling

Ewa Nowakowska
EY
Joseph Retzer
ACT Market Research Solutions

> *"What is the meaning of it, Watson? . . . It must tend to some end,or else our universe is ruled by chance, which is unthinkable" Sherlock Holmes*
> *The Adventure of the Cardboard Box*

Abstract

Traditional data clustering algorithms are challenged both by conceptual as well as practical issues. This paper will present biclustering, an approach which addresses the presence of dyadic relationships when clustering data. It will also profile the resultant "cluster cells" graphically with indicators of variable importance.

1. Introduction

1.1 Market Segmentation

As most market researchers are aware, market segmentation divides a market into smaller groups of customers with distinct needs, characteristics or behaviors who might require separate products or marketing mixes (Kotler and Armstrong, 2013). This allows managers to focus marketing campaigns, employee training, direct marketing efforts, etc., on specific customer groups for maximal effect.

Lilien et al. notes that "Market Segmentation is essential for marketing success: the most successful firms drive their businesses based on segmentation" (Lilien and Rangaswamy, 2004). This is testament to the value and importance of market segmentation analysis.

1.2 Challenges

With the advent of increased data availability, consumer information metrics have increased dramatically in both dimensionality (number of variables/features) as well as the number of observations (customers).

Consequences of clustering high dimensional data with traditional methods, and insufficient observations, include:

- The relative difference of distances between different points decreases with increasing dimensionality.
- Hence distances between points are less effective in differentiating between data point (customers).

- In fact, as dimensionality increases, the data distribution tends to degenerate to random noise.

In other words, without a sufficient number of observations, high dimensional data points tend to become almost equidistant from each other and therefore there are no clusters to discover.

The reason distances behave as they do in high dimensional spaces is due to the fact that typical distance functions give equal weight to all dimensions. All dimensions are not of equal importance however. It has been shown that adding irrelevant dimensions damages any clustering based on a distance function that equally weights all dimensions (See Kriegel et al., 2008). Indeed, in very high dimensions it is common for all objects in a dataset to be nearly equidistant from each other, completely masking underlying clusters.

The challenge of high dimensionality has been dealt with to varying degrees of effectiveness in supervised learning through techniques broadly referred to as "regularization" methods (e.g., L1, L2 and elastic net regularization). Market researchers, however, are often less aware of the challenges associated with high dimensionality when performing unsupervised learning.

Specifically, as dimensionality increases, the volume of the data space also increases dramatically making reliable identification of patterns/clusters increasingly difficult without sufficient data. Just how much data is sufficient for a given dimensionality was suggested by Formann in his 1984 paper (see Formann, 1984). Note that Formann was performing latent class clustering, however his suggestion was not necessarily specific to that method.

The recommended number of observations necessary to adequately identify useful clusters, assuming they exist, when there are p features is 5×2^p. So e.g., for p = 5, 10 or 30, the suggested sample sizes would be:

- $p = 5$, $n = 160$
- $p = 10$, $n = 5{,}120$
- $p = 30$, $n = 5{,}368{,}709{,}120$

Clearly, even with what may be considered as a typical sized feature set (p=30), required sample size is extraordinarily large.

The consequence of not having sufficient data is that respondents appear equidistant and are therefore randomly split resulting in a partition of low:

- quality,
- reproducibility, and
- predictive accuracy.

1.3 Dealing with High Dimensionality

Traditional data segmentation algorithms involve "one-way" clustering methods which form partitions by creating homogeneous groups of rows employing "all" columns/features.

When performing high dimensional data clustering, these approaches often lead to practical issues. This is due both to the expansion of the data space (as noted previously) as well as the presence of numerous irrelevant features masking existing clusters in noisy data.

1.4 Unsupervised Feature Selection

The challenge of high dimensional data clustering has been addressed in a variety of ways including what may be described as, "unsupervised feature selection."

Two such methods, which may be viewed as a natural progression of increasing generality, include:

- **Feature selection:** Feature selection assumes high dimensional data may contain many dimensions that are in fact irrelevant (and can mask existing clusters in noisy data). Feature selection first scores all data dimensions (features) based on various criteria, such as variance, entropy, ability to preserve local similarity, etc. (see Ciotan, 2019). It then identifies a subset of "relevant features/columns" which are used for clustering "all" rows/respondents in the data set. In other words, "all" individuals cluster on the same subset of features.[1]
- **COSA (Friedman and Meulman, 2004):** COSA (Clustering Objects on Subsets of Attributes) produces clusters which are defined by different subsets of features. COSA produces a distance matrix that serves as input for proximity analysis methods. This distance matrix contains the distances between "N" objects. The data set is assumed to have an underlying clustering structure in which the objects are clustered on cluster-specific subsets of attributes. To ensure this clustering is represented in the distance matrix, the attribute distances, for different clusters of respondents, are weighted using the "K Nearest Neighbors" (KNN) approach. It should also be noted that COSA produces clusters in which rows/respondents may only belong to a single cluster.

1.5 Factor-Cluster Analysis

Another popular approach to dealing with high dimensional data is "Factor-Cluster" analysis (often performed using Principal Component Analysis (PCA)). PCA is employed to create new basis variables which are weighted linear combinations of the original features (weights taken from data matrix eigenvectors). The number of principal components is usually determined by selecting only those with an associated eigenvalue greater than, or equal to, 1.

The original basis variables are then replaced with the principal components and the cluster analysis is performed.

[1] It should be noted that researchers will sometimes run Principal Component Analysis (PCA) prior to clustering and then select only variables that load strongly on each component/factor. This could technically also be considered a form of feature selection.

Some serious drawbacks to the Factor-Cluster approach however have been suggested by Dolnicar et al. (see Dolnicar and Grün, 2008). These include:

- The approach does not actually remove any of the original attributes from consideration, i.e., information from irrelevant dimensions is preserved.
- Averaging over the original features reduces variability making distinguishing clusters by definition more difficult.
- The data is transformed and segments are identified based on the transformed space, not the original information respondents gave which in turn may lead to different results. In addition, interpretations of segments based on the original variables is questionable given that the segments have been constructed in the space of the factor scores.
- Since typical explained variance is between 50 and 60 percent, up to half of the information that was collected from respondents is discarded before segments are identified or constructed.
- Eliminating variables which do not load highly on factors with an eigenvalue of more than 1 means that potentially the most important pieces of information for the identification of niche segments are discarded making it impossible to ever identify such groups.
- PCA may be best suited to datasets where most of the dimensions are relevant to the clustering task, but many are highly correlated or redundant.

2. Biclustering

2.2 The Model

While unsupervised feature selection models may aid in dealing with high dimensional data, an even more fundamental issue remains. Specifically, much of marketing data is comprised of “dyadic” (binary/pair-wise) relationships between two entities of interest. This in turn suggests clusters of rows (respondents) may be defined with respect to varying subsets of columns (features). Examples include “purchasing goods” (customer-product relation) or “brand appeal” (customer-brand relation) consideration sets.

An approach which may be used to uncover these dyadic relationships is known as “biclustering.” Biclustering simultaneously clusters on features and respondents producing subsets of individuals that cluster on different subsets of features. (Note: respondents maybe in more than 1 cluster.)

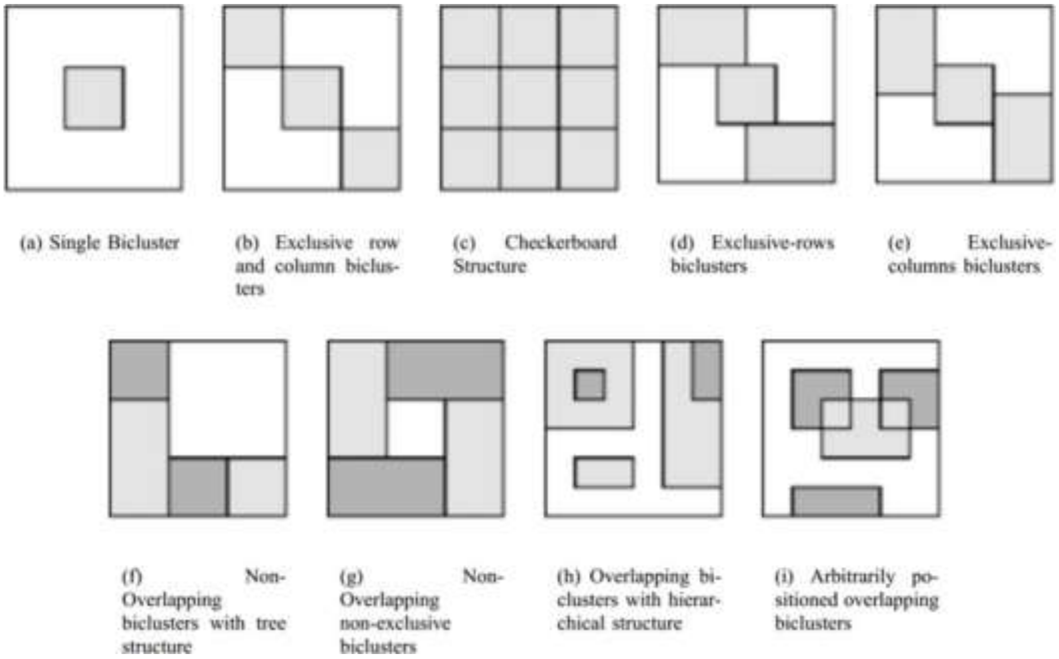
Biclustering was originally suggested by Hartigan in 1972 (Hartigan, 1972) but has only recently become popular due to its application to gene expression data (see Cheng and Church, 2000).

The biclustering approach is currently applied in many areas including market research. Biclustering (aka co-clustering) simultaneously clusters both rows and columns of the data set identifying homogeneous “cells” of row/column combinations. These cells are often visualized by an ordered heat map depiction of the solution.

It is important to note that biclustering is not a specific algorithm but rather an approach used to identify a two-way partition of data. This can be accomplished by a variety of algorithms that differ in:

1. types of data they deal with (e.g., binary, ordinal, metric) and
2. the “structure” of the bicluster solution (Wang et al., 2016).

Graphical illustrations of possible bicluster structures are shown in the figure below.



(a) Single Bicluster

(b) Exclusive row and column biclusters

(c) Checkerboard Structure

(d) Exclusive-rows biclusters

(e) Exclusive-columns biclusters

(f) Non-Overlapping biclusters with tree structure

(g) Non-Overlapping non-exclusive biclusters

(h) Overlapping biclusters with hierarchical structure

(i) Arbitrarily positioned overlapping biclusters

For this paper we selected “exclusive row biclustering” (d) in order to ensure each individual (row) would be a member of a single bicluster. Columns on the other hand, may participate in one or more clusters.

Reasons for choosing this structure are as follows:

1. The simpler structure facilitates communicating results to clients and makes marketing strategy easier to formulate.
2. In addition, it helps prevent smaller insignificant clusters from being formed.

An algorithm therefore needs to be selected such that it

1. is appropriate for the given data type (which is binary, as will be shown in section 3), and
2. ensures “exclusive row biclustering.”

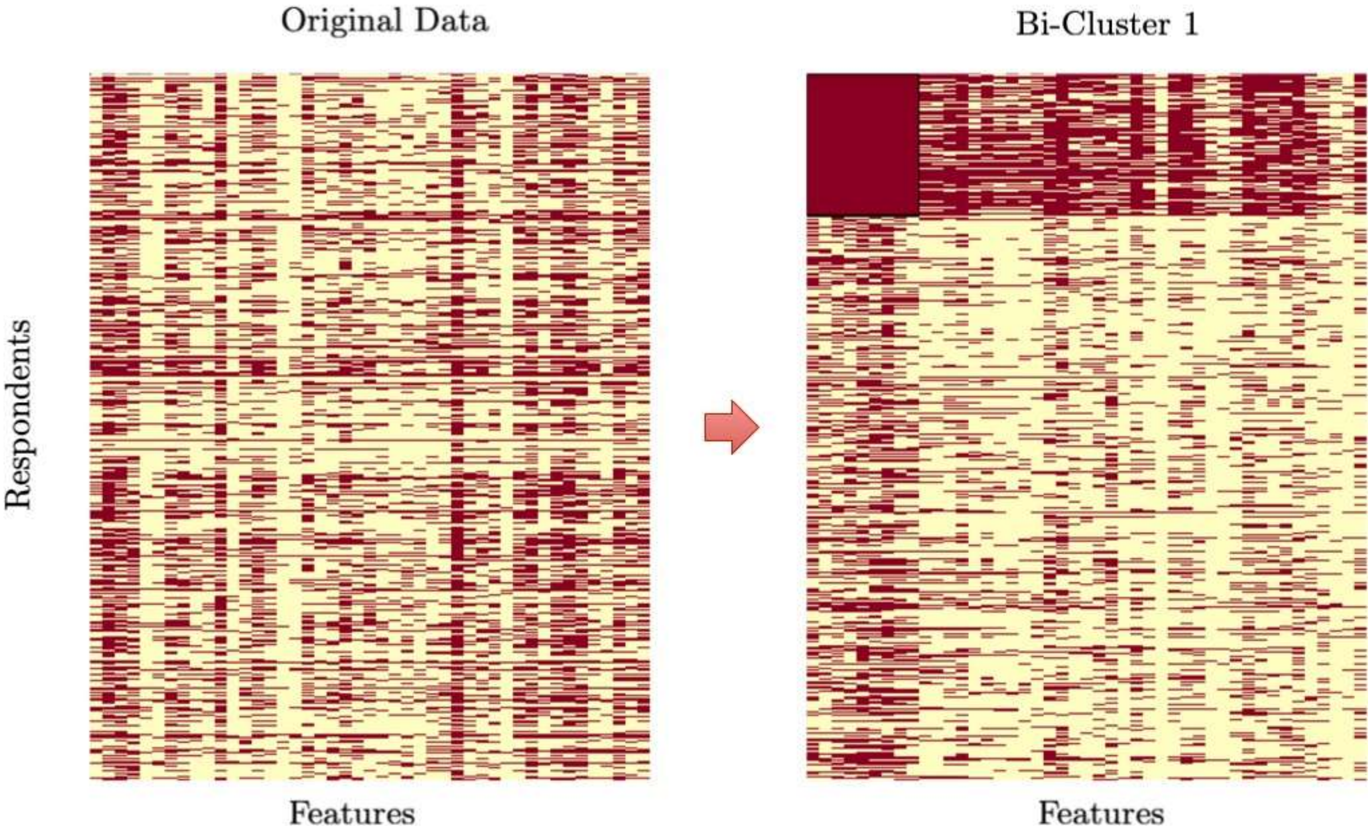
2.2 BCBimax

The algorithm chosen was "BCBimax," a modification of the Bimax algorithm suggested by Prelić (see Prelić et al., 2006). The Bimax algorithm was originally applied to gene expression data. It searches for sub-matrices of "1"s in a logical matrix using a fast divide-and-conquer algorithm.

BCBimax (Dolnicar et al., 2012) ensures an "exclusive row clustering" solution. BCBimax proceeds as follows:

1. Re-arrange rows (respondents)/columns (features) in order to find a sub-matrix of all 1's (1="yes" in this data). The first iteration of this step is illustrated in the heatmaps below.
2. Next, remove all rows in involved in the previous cluster and repeat step (1) to find the next cluster.

The process continues until the pre-specified number of clusters is reached. Note thatnot all observations are necessarily placed in a cluster.



3. The Data

Our data is taken from a survey of "New Auto Buyers." Respondents were asked what reasons for purchase were *extremely important* to them (binary response), for example:

- Overall interior styling,

- Riding comfort,
- Quietness,
- Passenger seating capacity, etc.

45 reasons (features) in total were used for the analysis.

4. Reproducibility

While a variety of metrics may be used to select an optimal cluster solution, we chose to focus on cluster reproducibility as detailed by Ernst and Dolnicar (2018).

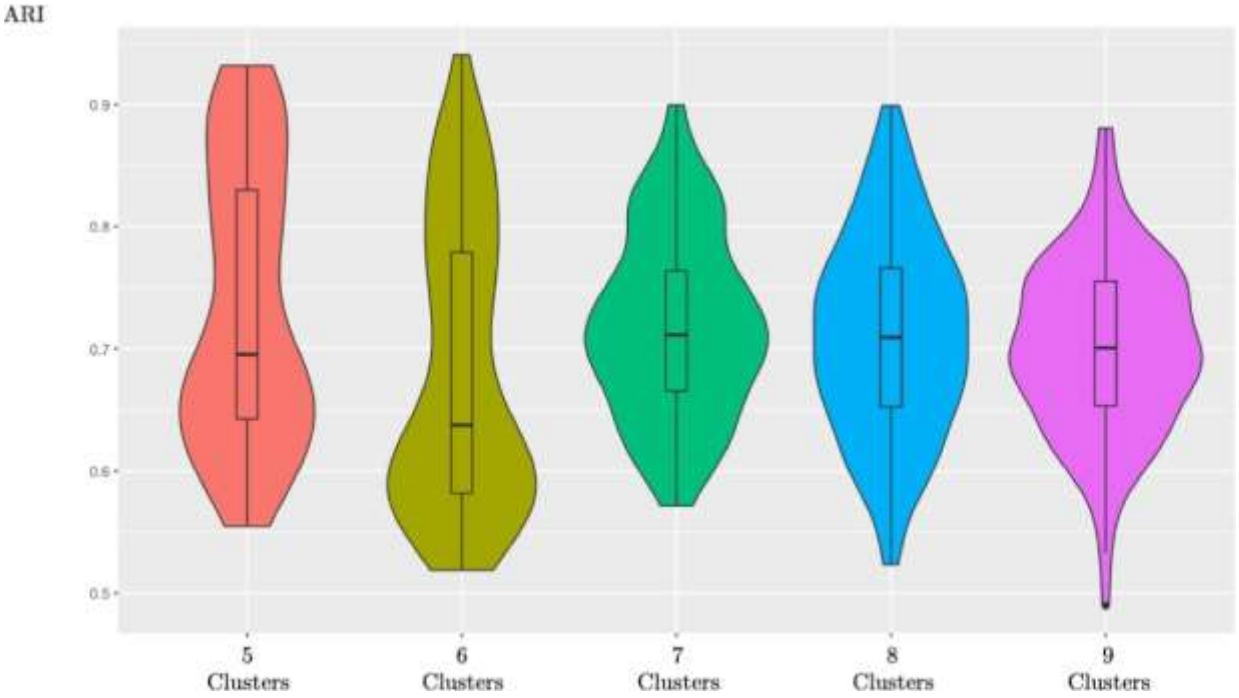
The process for evaluating reproducibility proceeds as follows:

1. Draw 2 bootstrap samples from the data and cluster both using some algorithm "*A*" and specified number of clusters "*k*."
2. Create 2 predictive models based on partitions from step (1) using, e.g., randomForest.
3. Predict the *original* data once with each predictive model from (2).
4. Determine the level of agreement between the predicted cluster membership vectors (3) using the "Adjusted Rand Index" (ARI).

(Repeat steps 1-3 a total of 200 times and create box-whisker plot.)

It should be noted that either the algorithm "*A*," or number of clusters "*k*" may be varied in order to select the optimal model. Here only "*k*" is being varied. We selected 5≤*k*≤9 as potential values for "*k*."

The box-whisker (violin) plots resulting from the reliability analysis are shown below.



Solutions $k = \{7, 8, 9\}$ all had maximum ARI values below the $k = \{5, 6\}$ solutions. In addition, each was concentrated (75th to 25th percentiles) between, roughly .76 to .67, as can be seen by the embedded box plots. Both the $k = \{5, 6\}$ solutions had higher 75th percentile scores (and higher max values) so those two were the solutions considered.

In comparing the $k = \{5, 6\}$ solutions it can be seen that while $k = 6$ had the maximum ARI, its violin plot clearly shows greater concentrations below .6. Therefore $k = 5$ was chosen as the optimal solution.

6. Results

6.1 Partition

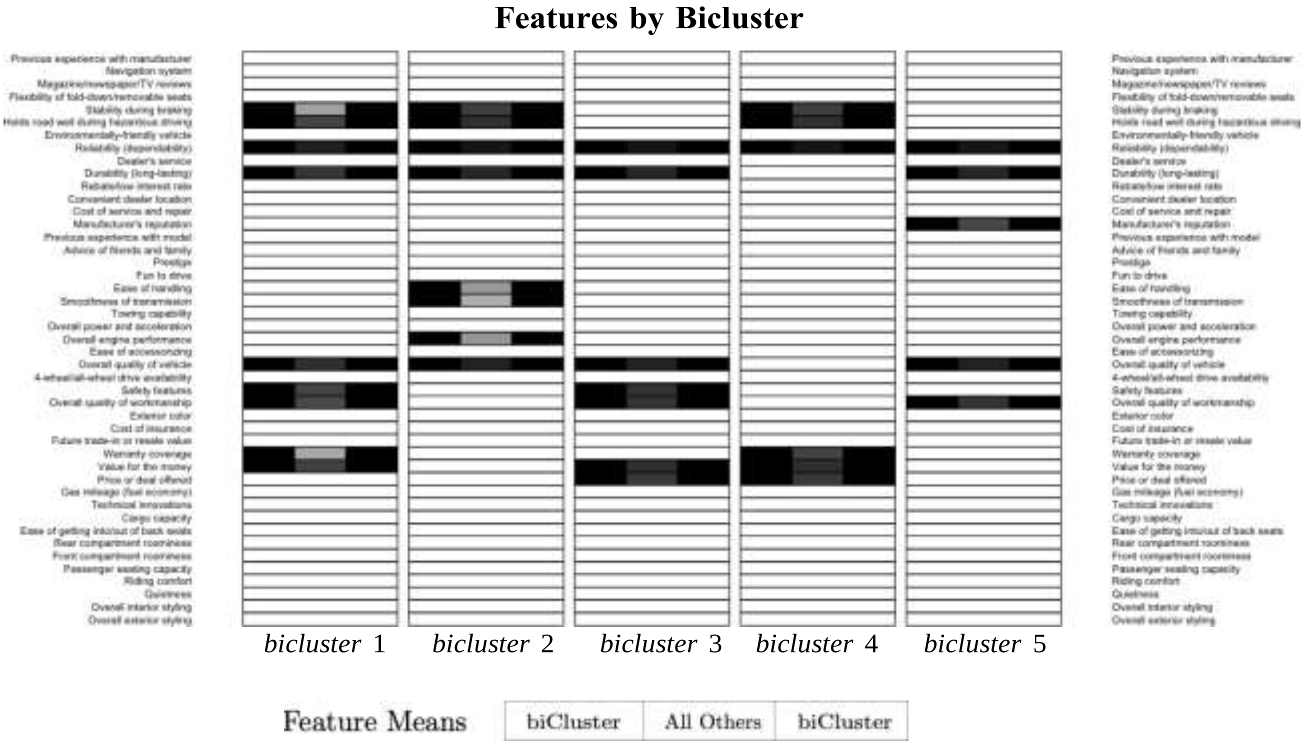
After applying the BCBimax algorithm to the data, the resultant rows/columns were produced as illustrated in the table below:

Bicluster Partition Results

	BC 1	BC 2	BC 3	BC 4	BC 5
Number of Rows / Respondents:	1186	284	302	258	300
Number of Columns / Features:	9	8	7	6	5

As can be seen, bicluster 1 has by far the most respondents as well as the most associated features.

In order to gain additional intuition into the nature of the biclusters, visualization using the `biclustmember` function found in the R `biclust` package was performed. The graphic produced by the function is shown below.



In the above graphic, each cluster is represented by a single column. An empty column "slot" indicates that the specific feature is not involved in the corresponding bicluster. A non-empty slot is shaded so as to indicate the mean of the feature. The darker the shading, the closer the mean is to 1.

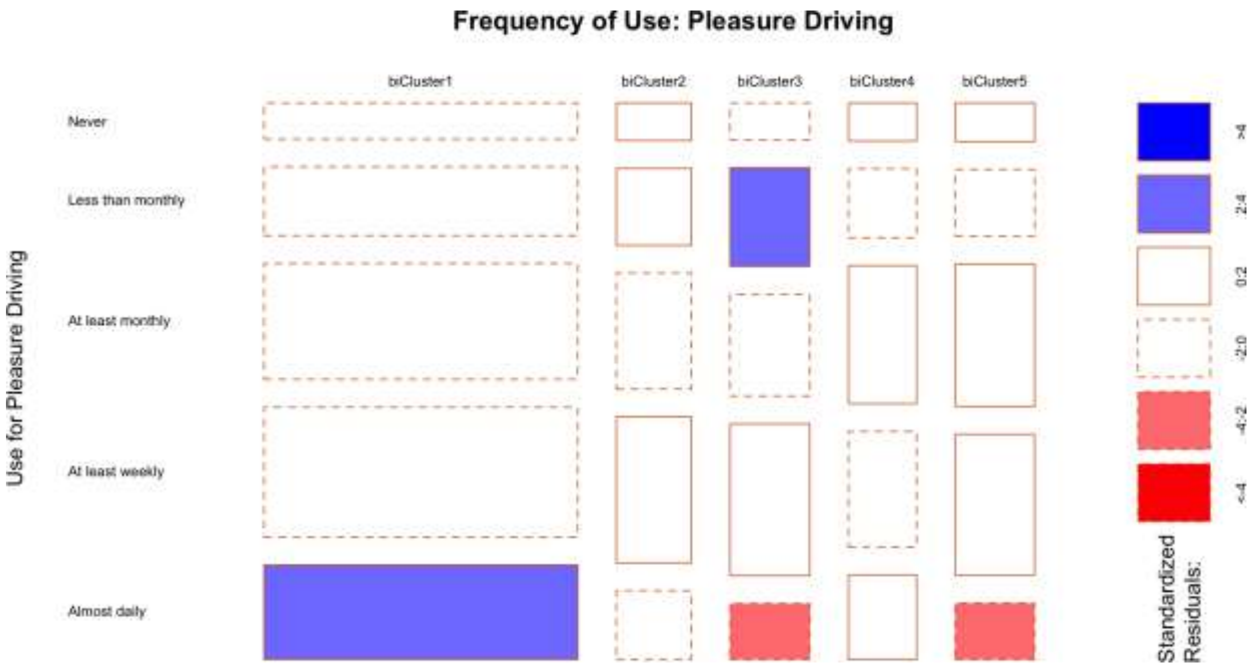
Note that the slots are divided into 3 sections. The middle section represents the feature mean in all biclusters other than the given bicluster. The left and right sections reflect the mean for the given bicluster.

So, for example, "Smoothness of transmission" is strongly associated with bicluster 2. In addition, it appears to be a unique feature for that bicluster as the center segment is relatively light in color.

6.2 Profiling

Since most profiling variables were either categorical or ordinal, an appropriate method for profiling such measures was needed. Graphical illustration via an extended Mosaic Plot was chosen to visualize the relationship between profiling variables and bicluster membership.

For the sake of illustration, we only present a single plot, for the feature "Use for Pleasure Driving."



* Two-way table depicted by rectangles where areas are proportional to cell frequencies.

In the mosaic plot above, the area of each rectangle represents the associated relative frequency. Note, for example, that the widths of bicluster 1 rectangles are relatively large compared to those of biclusters 2 through 5. This results in comparatively large relative frequencies for all levels of "Use for Pleasure Driving" associated with bicluster 1.

The mosaic plot is "extended" by color coding of the rectangles reflecting the significance associated with tests of independence between the level and bicluster. Again, for example, we see "Use for Pleasure Driving" being "Almost Daily" is significantly higher that would be expected under the assumption of independence. Alternatively, the level "Almost Daily" appears significantly lower than expected for biclusters 3 and 5 again under the assumption of independence.

7. Conclusions

Conclusions drawn from the analysis may be summarized by listing both the advantages as well as caveats relating to biclustering.

Advantages

Biclustering addresses the challenge of high dimensionality by automatically selecting subsets of features *without transformation*. Unlike unsupervised feature selection, the subsets are selected while taking into account potential dyadic relationships between features and respondents. Avoiding space altering transformations (as in factor-cluster analysis) is also advantageous in that it circumvents issues resulting from this approach (see section 1.5).

- Various algorithms are available to deal with specific data types allowing the approach to be employed for a variety of feature sets.
- Respondents may be classified in multiple partitions if so desired. While this was deliberately avoided in this study, it represents a unique and potentially informative extension of typical clustering results.
- Biclustering partitions have high predictive accuracy. While predictive accuracy was not reported, the biclustering partitions did in fact, and are generally known to, produce highly accurate typing tools.

Caveats

- Computational demands may result in long run times. This challenge was in fact experienced in this study particularly as the number of desired clusters specified was increased.
- Biclustering (BCBimax) doesn't necessarily classify all observations. While this may be disconcerting for some clients, it in no way limits subsequent typing tools from providing accurate predicted cluster memberships for any/all potential future customers.

Acknowledgments

The authors wish to thank Dr. John Reisner, Department of Statistics, Iowa State University, Dr. Friedrich Leisch, University of Natural Resources and Life Sciences, Vienna, and Dr. Sebastian Kaiser, University of Duisburg-Essen for their valuable advice and assistance in the preparation of this presentation.

Ewa Nowakowska

Joseph Retzer

References

Cheng, Y. and Church, G. M. (2000). Biclustering of expression data. In *Ismb*, volume 8, pages 93–103.

Ciotan, M. (2019). Overview of feature selection methods. *Bioinformatics*.

Collins, L. M. and Dent, C. W. (1988). Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research*, 23(2):231–242.

Dolnicar, S. and Gru¨n, B. (2008). Challenging “factor–cluster segmentation.” *Journal of Travel Research*, 47(1):63–71.

Dolnicar, S., Kaiser, S., Lazarevski, K., and Leisch, F. (2012). Biclustering: Overcoming data dimensionality problems in market segmentation. *Journalof Travel Research*, 51(1):41–49.

Ernst, D. and Dolnicar, S. (2018). How to avoid random market segmentation solutions. *Journal of Travel Research*, 57(1):69–82.

Ewoud, D. T. (2020). The biclustgui r package vignette - 1.1.3. *TheComprehensive R Archive Network*.

Formann, A. (1984). Die latent-class-analyse: Einführung in die theorie und anwendung.

Friedman, J. H. and Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society:Series B (Statistical Methodology)*, 66(4):815–849.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Kotler, P. and Armstrong, G. (2013). Principles of marketing (16th global edition).

Kriegel, H.-P., Kr¨oger, P., and Zimek, A. (2008). Detecting clusters in moderate-to-high dimensional data: Subspace clustering, pattern-based clustering, and correlation clustering. *Proc. VLDB Endow.*, 1(2):1528–1529.

Lilien, G. L. and Rangaswamy, A. (2004). *Marketingengineering: computer-assisted marketing analysis and planning*. DecisionPro.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

Wang, B., Miao, Y., Zhao, H., Jin, J., and Chen, Y. (2016). A biclustering-based method for market segmentation using customer pain points. *Engineering Applications of Artificial Intelligence*, 47:101–109.

CONJOINT FOR DIFFICULT CHOICES

JOEL HUBER
DUKE UNIVERSITY
MARTIN MEIßNER
ZEPPELIN UNIVERSITY

ABSTRACT

Choice-based conjoint has successfully predicted choices across many domains of decision-making. This paper explores what to do when that prediction is hindered because the conjoint choices are difficult for the respondent. Decisions are difficult when the choice is novel and requires learning, when the attributes are themselves hard to understand, and when trade-offs are unsuitable because they include unacceptable or irreversible outcomes. We provide specific recommendations to deal with such difficult decisions. To increase the motivation to make conjoint choices and support their relevance in the marketplace, it is important to select respondents who are involved in the decision. Unfamiliar attributes can be clarified by focusing attention and emotion on a limited number of levels and encouraging thoughts for and against each level. Unwillingness to make trade-offs can be alleviated by using paired comparisons on preference rather than absolute choice. We show that such careful preparation can help us clarify decision-making for difficult problems.

INTRODUCTION

Choice experiments assess the trade-offs people are willing to make between alternatives. They work well in contexts where respondents are familiar with the relative value of the attributes. Thus, for products often consumed it is relatively easy to decide between a heavy detergent box with a low price per pound against a more convenient size with a higher price. However, conjoint's accuracy may be limited when respondents are unfamiliar with the attributes or have not thought about how to trade them off against each other. Paradoxically, the results of the choice-based conjoint in such cases may appear correct, indicated by measures of fit or even respondent perceptions. However, the ability of people to make conjoint choices that reflect what they would do can be severely limited if respondents are not prepared for appropriate decision contexts. Our goal here is to suggest ways to select respondents, teach them about the attributes, and simplify the choice sets to generate effective conjoint exercises.

In the next section we provide ways to motivate respondents to make difficult decisions. We consider the selection of relevant respondents and the framing of the choice task, ways to encourage thoughts about attribute levels, and ways to facilitate choice evaluations by simplifying the choice structure and labeling of attribute levels. Following that we focus on dealing with a particularly problematic attribute: probabilistic outcomes. We summarize what is largely accepted when communicating probabilities and give an example of appropriate use of probabilities in a conjoint study. Finally, we discuss the outcome of a study investigating a decision about an important health procedure that turned out to be inappropriate for conjoint, and we outline an alternative procedure that encourages deeper understanding of that decision.

Three Ways to Encourage Respondents to Make Difficult Choices

Select appropriate respondents. The easiest way to facilitate careful and thoughtful responses is to recruit respondents who already care about the decision. When studying the strongly growing market for electric vehicles, current brand users or those known to be in the car market are good subjects. They may not have thought through the attributes (such as charging times), but since they are automobile users, they are motivated to carefully consider the decisions. It is possible to ask a person to imagine they need a particular product or service, but if that prospect is unlikely or unfamiliar, the responses may reflect what the respondent rightly or wrongly projects to others.

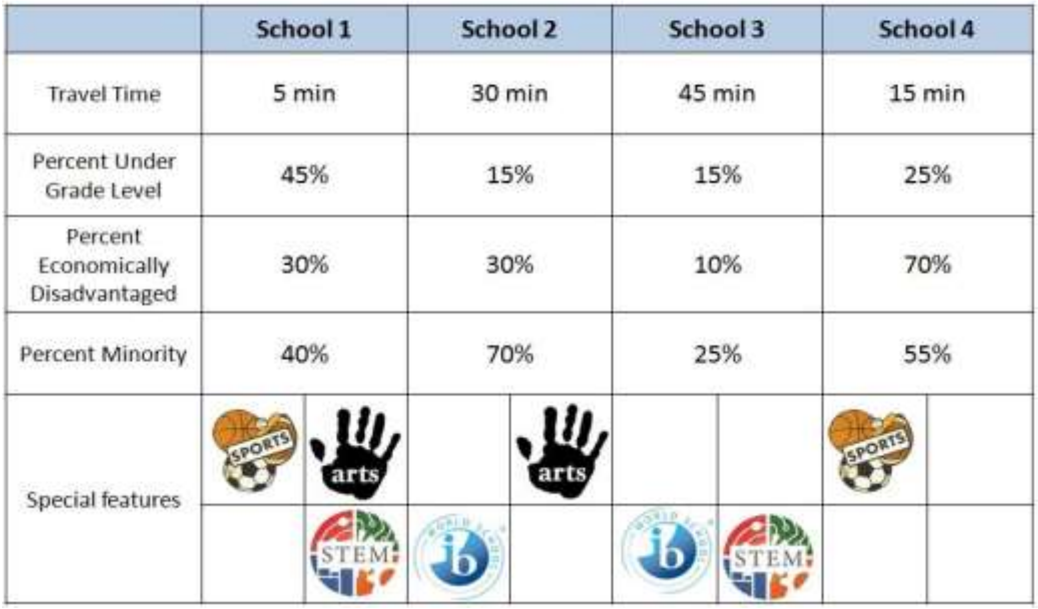
Carefully introduce attributes. Given a respondent is involved with the choice it is important to carefully introduce each attribute level. As each level is introduced, it is helpful to encourage thinking about the problem. For example, when describing an automotive feature enabling automated parallel parking, consider asking whether the respondent had tried to parallel park in the last month, and if so, how difficult it was. At times in our research we ask direct importance questions (similar to the direct attribute evaluations in Adaptive Conjoint Analysis). These preliminary ratings are less important in estimating conjoint utilities but instead encourage stable choices among alternatives with conflicting features.

Limit the number of attributes and levels. For difficult choices, it is helpful to limit the number of attributes and levels. Describing choice alternatives with too many features makes it very difficult for respondents to generate consistent trade-offs. For continuous attributes and particularly price, choices are facilitated if the number of levels is reduced to three or four. Leaving a continuous level unbounded can generate inconsistent choices. For example, prices can be rejected if they are too low or too high. Instead, it is better to choose a few levels that reflect a range expected by the respondent. From a respondent's perspective it is far easier to make decisions on options with three prices, one being about right, the second one being relatively high, and the third being relatively low.

An Illustrative Example: School Choice

These ways to deal with complex decisions are apparent in a study presented at Sawtooth Software's 2013 conference that estimated parents' preferences for high schools (Fairchild, Sagara and Huber 2013). The conjoint tasks involved eight choice sets similar to the example shown in Figure 1. The task is both complex and difficult, involving identifying the best and worst of four profiles each defined by eight attributes. The study design illustrates ways to reduce the difficulty and increase the meaningfulness of the task with appropriate selection of respondents, a careful introduction to the choice task, and thorough presentation of the attributes and levels.

Figure 1: A Best-Worst Conjoint for High School Choice



	School 1	School 2	School 3	School 4
Travel Time	5 min	30 min	45 min	15 min
Percent Under Grade Level	45%	15%	15%	25%
Percent Economically Disadvantaged	30%	30%	10%	70%
Percent Minority	40%	70%	25%	55%
Special features	SPORTS arts STEM	arts ib	ib STEM	SPORTS

The respondents were panel members predefined as having at least one child age 11 to 17 years in public school. School choice is increasingly popular in the United States as districts seek to enable parents to select schools that match their goals for their children. To frame the problem, we told respondents "Suppose you just moved to a new area where families are able to choose which school they would most like their children to attend."

The survey began by defining each attribute and relating its levels to the respondent's experience. For example, after showing the four levels of 15% to 45% percent under grade level, it asked "At your current school, what percent are below grade level?" The other four-level attributes were bus or car travel time, percent economically disadvantaged measured by the fraction of students on free lunch, the percent under grade level, and the percent of minority students. Additionally, the four binary variables reflect areas of strength in the school: Sports, Arts, STEM (science, technology, engineering, and mathematics), and IB (International Baccalaureate).

Defining levels with simple labels and identifiable images facilitated the respondent's job of selecting the best alternative in the set. Additionally, once the best alternative is identified, then finding the worst is easier because features have already been considered by the respondent. The combined best-worst choices provided sufficient information to generate stable utilities for each respondent. Among other findings, the analysis of the part-worth values showed that parents with higher levels of education strongly preferred schools with high academic quality while parents with less education desired schools with active sports and arts programs.

The use of Best-Worst Scaling (MaxDiff) for conjoint profiles is relatively rare. Most choice experiments focus on finding the best product, service, or experience. That makes sense for goods where there are a large number of reasonable alternatives. In those cases, there is relatively little value in knowing which recipe, trip, or brand of toothpaste is *least* liked. However, for school choice it is important to identify deeply upsetting school characteristics.

The focus on factors to accept and avoid is most appropriate for long-term decisions where any option includes both positive and negative aspects, and the school choice example illustrates the value of conjoint for this kind of decision. Before making conjoint choices, respondents were encouraged to think about the value of attribute levels defined by consistent labels or easily recognizable images. Thus, while eight versions of the task similar to those shown in Figure 2 may seem daunting, framing a relevant topic and simplifying the presentation of attributes and levels helped generate high-quality preference measurement.

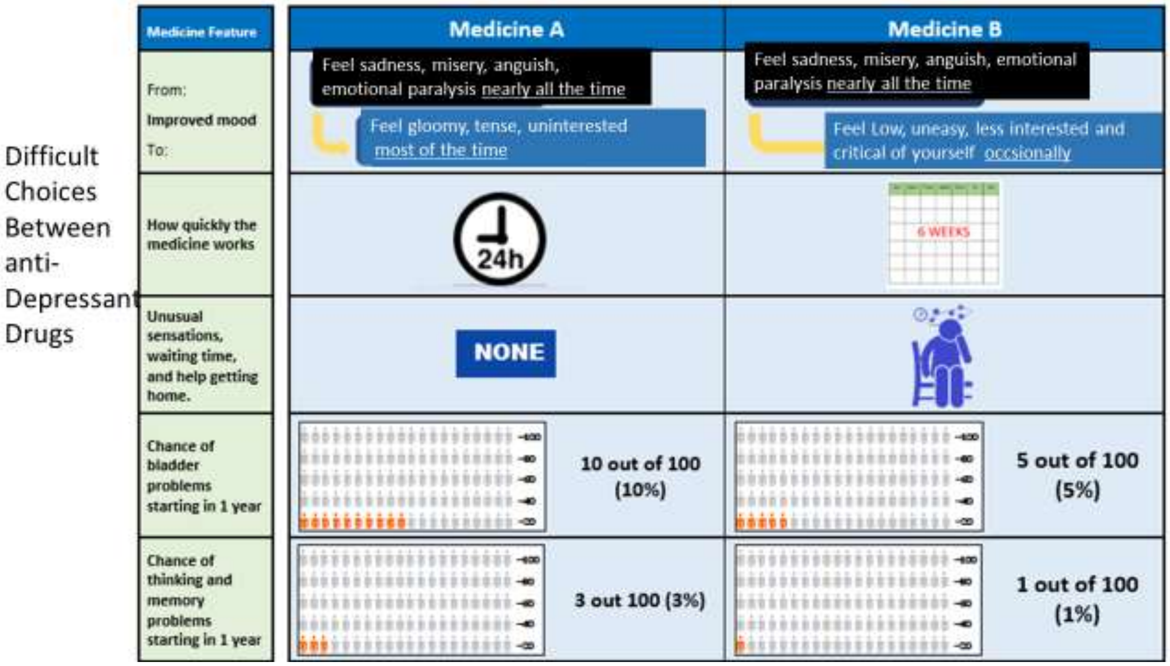
What to Do About Probabilistic Conjoint Attributes?

Probabilistic outcomes can increase the difficulty of any decision. Bonner et al. (2021) wrote a fine review of the appropriate ways to present probabilities to patients concerned with the risks and benefits of medical decisions. Below are some conclusions from that review that apply generally to probabilities in conjoint choices.

1. **People differ strongly in their ability to handle probabilities.** People with limited experience dealing with probabilities may require probabilities presented in different ways. Respondents can be grouped by a simple ability test. For those failing the test, it may be useful to include visual representations of risk (as shown in Figure 2) or more intuitive subjective probabilistic expressions like “very good chance.”
2. **People generally have difficulty understanding small or large probabilities.** Probabilities are best used when greater than .01 and less than .99. More extreme probabilities may be transformed to an acceptable range by assembling risks across time or populations. For example, a very low one-year by probability of harm can be increased by assessing a 10-year period or the harm for a town of 100,000. Whenever such aggregation is used, it is important that the same frame be used across competing probabilistic attributes.
3. **Probabilities work best if expressed in terms of a count out of a base number.** Odds ratios and proportions are generally harder to understand.
4. **Visual count images are better ways to show numeric risks instead of length or area representations.** Showing the number of cases as points or faces in a 20 x 5 grid provides relative risk and absolute risk precision if needed.
5. **Use redundant measures of probabilities.** Provide counts out of a base, percentiles, and graphs to represent the same probability multiple ways.

The recent article by Fairchild et al. (2021) illustrates these principles in a choice-based conjoint study. It examines patient response to a novel anti-depression drug that may cause bladder or memory problems. Figure 2 provides an example of their conjoint task where the focal question involves the choice between two medicines that differ on reduction of depression and probabilities of harm. The top rows show the degree of improvement. Those benefits are balanced against speed of action, a minor short-term dizziness, and the probabilities of bladder or memory deterioration a year after taking the drug.

Figure 2: Risk Probabilities in Choice-Based Conjoint an Anti-Depression Drug



Difficult Choices Between anti-Depressant Drugs

Medicine Feature	Medicine A	Medicine B
From: Improved mood To:	Feel sadness, misery, anguish, emotional paralysis nearly all the time Feel gloomy, tense, uninterested most of the time	Feel sadness, misery, anguish, emotional paralysis nearly all the time Feel Low, uneasy, less interested and critical of yourself occsionally
How quickly the medicine works	24h	6 WEEKS
Unusual sensations, waiting time, and help getting home.	NONE	
Chance of bladder problems starting in 1 year	10 out of 100 (10%)	5 out of 100 (5%)
Chance of thinking and memory problems starting in 1 year	3 out 100 (3%)	1 out of 100 (1%)

Notice that the probabilities have redundant labels defining a risk including the number inflicted out of 100, then as a percent, and finally with the number of orange figures in a 20 x 5 grid. The idea is to allow respondents to understand the risk in a way that is most meaningful. Further, note that both the one-year time period and the probability denominators are equivalent across the two outcomes. Finally, the visual framing allows the processing of either absolute or relative risks. In the example given, medicine B has twice the bladder cancer risk in absolute terms but three times the risk of memory loss of medicine A. However, memory loss has a far lower likelihood of occurring. In all then, the design demonstrates reasonable ways to include probabilities in difficult but important decisions.

There are several other learnings that can be derived from this study. Having only two alternatives simplifies the choice for this new type of drug. The study is less about choice among alternatives, but whether patients with drug resistant depression are willing to accept risks to get significant mood improvements. The authors noted that around 40% of respondents focused on one attribute, implying an unwillingness to trade off gains in one attribute against all others. Many conjoint studies try to discourage this natural unwillingness to avoid trade-offs. However, in this case identifying the maximum probability of harm that leads a severely depressed person to accept a less effective medicine is itself useful for public health policy and is one of the key contributions from this paper.

When Classic Conjoint Can Go Wrong

Conjoint is so reliable that its failure can be surprising. We were working to help surgeons identify appropriate characteristics of candidates for human hand transplants. Hand transplants are a promising alternative to prosthetic appliances. The cost of a transplant is great but currently subsidized. It was important for surgeons to understand how potential patients might react to various medical options if they lost a hand.

We began with a single conjoint exercise which asked MTurk respondents to assume that they had lost a hand and needed to consider options for the future. They were shown the attributes for prosthetics and separately the attributes of a transplant process. The four attributes for the prosthetic hand were appearance (mechanical, lifelike, hybrid), function (passive, body link, myoelectric), maintenance (free, regular, replace), and training time (1, 3, or 6 months). For the transplant the attributes were function (minimal, light, full), maintenance (local, hospital, replace), grasp (minimal, light, full functions), and the years of required immune suppressant drugs (5 years, 10 years, life). For each domain we introduced the attribute levels and asked respondents to indicate how they felt about each. They then rated 10 conjoint profiles for prosthetic and 10 profiles for the transplant hands. Following both exercises we asked respondents to select between a prosthetic and a transplant hand shown below:

Test Choice between a Prosthetic and Transplant Hand

Prosthetic Hand	**Transplant Hand**
Appearance: Hybrid	**Sensation:** Protective
Function: Body powered	**Grasp:** Light objects
Maintenance: Minor	**Maintenance:** Hospital
Learning time: 3 months	**Drugs:** 5 years

Our hope was to be able to use the separate conjoint scores for the prosthetic and transplant profiles to predict this choice. On the first run with 100 respondents, we realized that the choice between domains with different attributes was too difficult. Over 80% of respondents simply rejected the transplant. Further, neither conjoint scores nor respondent characteristics came close to reliably predicting choice.

Chastened by our unsuccessful start, the next version was specifically designed to provide physicians with information on potential patient willingness to begin an elaborate transplant process. For 400 new respondents we retained the conjoint on the prosthetic hand as a way to encourage thinking about the ways to conventionally deal with a lost hand. Then respondents engaged in much more detail on the transplant process and provided their reactions to positive and negative statements ordinary people make about such a procedure.

Below are positive statements organized by the percent valued among the top three. These results show that the integration of the hand with the recipient's body is more important than functional benefits of a transplant.

Positive Statements about Transplants Most Valued

Top 3	
72%	Its connection to my body is natural, just like my real hand
67%	Feeling and strength will gradually develop with time
51%	A transplanted hand would be part of me
46%	I train the transplanted hand with exercise and work
45%	People can hold my hand, and shake my warm transplanted hand
20%	My transplanted hand is cleaned and manicured like my other hand

The next tasks asked respondents to indicate their level of concern reflected in negative statements.

Negative Statements about Transplants of Most Concern

Top 3	
68%	Requires immune suppression drugs
58%	Intensive surgery
38%	Limited likely grasp
37%	Limited likely sensation
31%	Having to accept a hand from another person
28%	Visual difference in your transplanted hand
26%	Extensive physical therapy

These selections demonstrate that respondents are mostly worried about the immune suppressive drugs and the extent of surgery required. These evaluative tasks are not intended to move the attitude in either direction, but to encourage awareness of the costs and benefits of the transplant process, and to prepare respondents for the following critical choice question:

The Critical Transplant Question

> Now suppose you had an artificial hand that served your purposes well, but you had an opportunity go through the process of being fitted for a transplant hand.
>
> There would be no out-of-pocket cost to you, but the selection, surgery, rehab, physical therapy, and limited abilities would be as described earlier.

Would you be likely to enter such a program?

Yes **No**

Notice the question asks whether they would be *likely* to enter such a program. That encouraged respondents to be open to revising their initial decision. We assessed that likelihood by providing statements by a physician offering information against the initial decision. Below we show the negative information provided by the doctor.

Questions Seeking Changes in Choice with New Information

You said you would be likely to enter a hand transplant program.

1. Suppose your doctor told you that given your health and the type of injury you suffered to your hand, the chance of you developing a fully functional transplanted hand would be less likely than most patients.

Would this possibility change your mind?

Yes No

Ask if ***No***

2. Suppose your doctor told you that due to certain characteristics of your immune system, you will be required to take strong immunosuppression drugs for the rest of your life. These drugs will have all the risks that were previously described including a risk for developing kidney problems, cancer, diabetes, and infections.

Would you still be likely to enter a hand transplant program?

Yes No

If a person held to their initial judgment against both counters then that indicates the initial commitment to a transplant is strong, and it gets a transplant support score of 3. If the first counter is resisted but not the second, it gets a score of 2. If the respondent reverses after the first counter, it gets a score of 1. The same process is reversed for those initially indicating they are *not likely* to enter the transplant program, except the counters give favorable rather than unfavorable transplant information. For example, the first counter may indicate that the chance of developing a fully functional hand would be more, rather than less likely than most patients. Then, just like the counters to a positive response, the degree of negative response is coded -1, -2, and -3, providing a reasonable continuous measure of support for transplants.

Figure 3: Percent Supporting Transplant Given Initial Judgment and Physician Counters

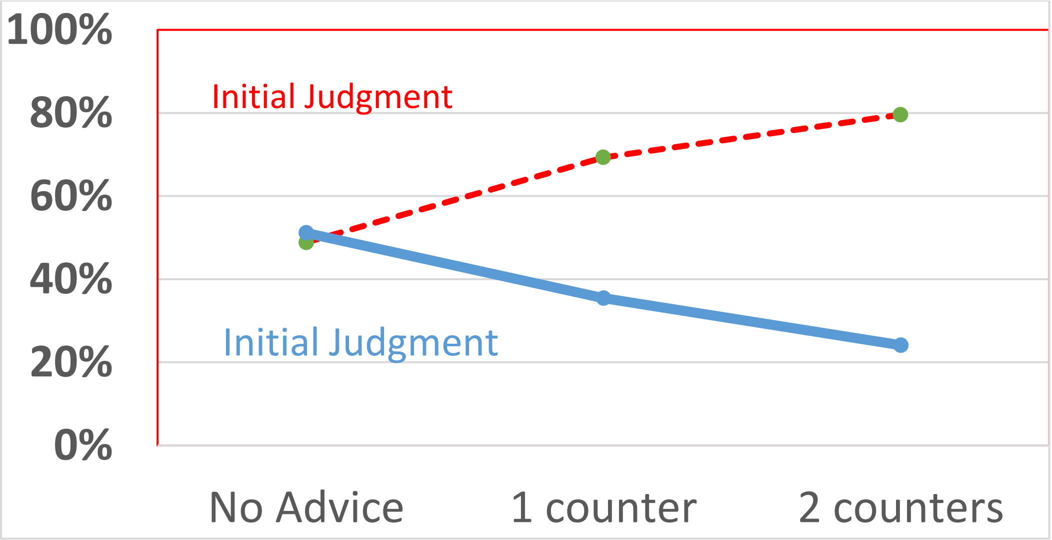


Figure 3 shows the impact of new information from physicians on the percent supporting a hand transplant. About 50% supported the transplant initially. However, the solid blue line shows that nearly 20% became negative after one and another 10% switched after the second counter. The red dotted line shows that for those initially against the transplant, the counters increased the likelihood of a positive decision by nearly the same amount. These shifts vividly demonstrate the importance of information from knowledgeable physicians on this decision.

The -3 to +3 scale also provides a richer measure that augmented our insight into the characteristics of those who support transplants. Below are the characteristics of respondents who have significantly stronger transplant support, with the strongest effects given first.

Transplant supporters are those who

- Are younger and supporters of technology
- Have more medical knowledge and are less worried about drugs
- Believe they are relatively healthy
- Attend regular religious services
- Had military service and are more likely to be male
- Continue risky behavior: heavy lifting, chain sawing
- Prefer lifelike over mechanical-looking artificial hands

The first three results make sense medically. Those who are healthy and comfortable with medical processes are good prospects for the operation. The last four make sense psychologically, suggesting that physicians could focus on recruiting candidates with military experience and strong religious support.

The transplant study also has value as a case study of what to do when standard choice-based conjoint does not work. Conjoint is most appropriate when there are multiple options each with clearly defined features. The decision to accept the prospect of a transplant is not a choice among many options, but reflects respondents' willingness to move towards a hopeful but uncertain outcome. Thus, the goal is to provide information about the cost and benefits of such a decision, encouraging a preliminary decision, and assessing the impact of information provided by a physician. It is then possible to determine the characteristics of those most likely to be part of a risky medical trial, and the reasons that are most effective, enabling potential patients to make an effective medical decision with the doctor.

Conclusion

Our primary goal has been to suggest ways to better prepare respondents for difficult choice tasks. The better respondents understand the attributes and the advantages and disadvantages of their levels, the more likely they will be willing to make reasonable trade-offs. Consistent choices are facilitated by limiting the number of distinct attribute levels and focusing on judgments among paired rather than multiple alternatives. Prediction to actual choice improves for choices from respondents who express a need for the product category and consider options across a limited domain. These recommendations may be less important for market-based conjoint exercises where the features are relatively well known and where there are multiple competitive offerings providing different features. However, we provide examples of school and health care decisions that require more elaborate preparation. Additionally, the decision to replace a prosthetic with a transplant human hand requires thinking about the advantages and disadvantages of a transplant hand in a context where commitment is gradual rather than absolute. We illustrate how to identify appropriate prospective transplant candidates and how to assess the impact of physician statements that can help patients make very difficult decisions.

Joel Huber

Martin Meißner

References

Bonner, Carissa, Trevena, Lyndal J., Gaissmaier, Wolfgang, Han, Paul K. J., Okan, Yasmina, Ozanne, Elissa, Peters, Ellen, Timmermanns, Danielle, & Zikmund-Fisher, Brian J. (2021). Current Best Practice for Presenting Probabilities in Patient Decision Aids: Fundamental Principles. *Medical Decision Making*, 0272989X21996328.

Fairchild, Angelyn O., Namika Sagara and Joel Huber (2013) Best-Worst CBC Conjoint Applied to School Choice: Separating Aspiration from Aversion. *Sawtooth Software Conference Proceedings* (2013) pp 317–329.

Fairchild, Angelyn O., Katz, E. G., Reed, S. D., Johnson, F. R., DiBernardo, A., Hough, D. and Levitan, B. (2020). Patient Preferences for Ketamine-based Antidepressant Treatments in Treatment-resistant Depression: Results from a Clinical Trial and Panel. *Neurology, Psychiatry and Brain Research*, 37, 67–78.

CONCORDANCE BETWEEN TREATMENT CHOICE AND PREFERENCE FOR LOCALIZED PROSTATE CANCER

RAVISHANKAR JAYADEVAPPA[1]
SUMEDHA CHHATRE[1]
JOSEPH J. GALLO[3]
MARSHA WITTINK[4]
KNASHAWN H. MORALES[1]
DAVID I. LEE[1]
THOMAS J. GUZZO[1]
NEHA VAPIWALA[1]
YU-NING WONG[2]
DIANE K. NEWMAN[1]
KEITH VAN ARSDALEN[2]
S. BRUCE MALKOWICZ[1]
ALAN J. WEIN[1]
[1] UNIVERSITY OF PENNSYLVANIA, PHILADELPHIA, PENNSYLVANIA
[2] CORPORAL MICHAEL J. CRESCENZ VA MEDICAL CENTER, PHILADELPHIA, PENNSYLVANIA
[3] JOHNS HOPKINS UNIVERSITY, BALTIMORE, MD
[4] UNIVERSITY OF ROCHESTER SCHOOL OF MEDICINE AND DENTISTRY, ROCHESTER, NY

ABSTRACT

Introduction and Objective: Treatment choice for localized prostate cancer is preference sensitive. The objective of our novel study was to assess the association between value markers (utility levels) and treatment choice in localized prostate cancer patients.

Methods: In this multi-center, longitudinal, randomized, controlled design study, localized prostate cancer patients were randomized either to a preference assessment intervention or to the usual care arm. Patient reported outcomes (satisfaction with care, satisfaction with decision, and generic and prostate-specific health related quality of life, depression and anxiety) were measured at baseline, and at 3, 6, 12 and 24-month follow-up. Clinical data such as stage of cancer, treatment, PSA, Gleason score and comorbidity were obtained from medical charts. Preference assessment was done using our web-based adaptive choice-based conjoint analysis tool, PreProCare, prior to treatment choice. For prostate cancer patients from the intervention group, a logistic regression model was used to analyze the association between value markers of treatment attributes and treatment, including active surveillance treatment.

Results: Between January 2014 and March 2015, 743 localized prostate cancer patients were recruited and randomized to the PreProCare intervention (n=371) or to a control group (n=372). The sample demographics and clinical characteristics were comparable by intervention status. Satisfaction with care showed significant improvement for the intervention group, compared to usual care. The scores on satisfaction with decision showed that satisfaction with decision improved in both groups. However, the improvement was significantly greater for the intervention group. Similarly, decision regret declined in both groups, however, the decline was greater for the intervention group. Treatment choice by prostate cancer risk category showed that in low-risk group, a higher proportion from the intervention group were on active surveillance compared to the usual care group (66% vs. 54%; p=0.24). Among intermediate-risk category patients, this proportion was 12% vs. 13% (p=0.76). Finally, 93% of the intervention group from the high-risk category were on active treatment, compared to 95% from the usual care group (p=0.72). Among 371 prostate cancer patients from the intervention group, attribute of survival was associated with higher odds of radiation treatment (OR=1.20, CI=1.06, 1.50), and surgery (OR=1.11, CI=1.06, 1.28). On the other hand, survival was associated with lower odds of being on active surveillance (OR=0.67, CI=0.48, 0.94). Sexual function attribute was associated with higher odds of being on active surveillance (OR=1.46, CI=1.04, 2.06). Finally, fear of cutting was associated with lower odds of surgery (OR=0.88, CI=0.77, 0.99) and higher odds of being on active surveillance (OR=1.54, CI=1.04, 2.24).

Conclusions: Preference assessment is a key component of patient-centered care and is feasible among localized prostate cancer patients. Results of our novel study showed that preference assessment was associated with improved satisfaction with care, satisfaction with decision and lower regret. Also, patient treatment choice aligned with their values. Value markers (or utility levels) of treatment such as survival, and fear of surgery were associated with active surveillance. Preference assessment intervention can help prostate cancer patients reveal their preferences, leading to better alignment with treatment decision. Future research should identify strategies to ensure diagnosis and treatment options are communicated to patients accurately, therefore reducing overtreatment and the resulting burden on healthcare systems.

INTRODUCTION

There is only limited evidence regarding the added benefit of value clarification methods through preference assessment in decision-aids to facilitate patient decision-making. There are several types of value clarification methods, and it is unclear what type is most helpful.[1,2] Few randomized controlled trials designed to assess the impact of enriching decision-aids with value clarification methods suggest they improve the match between values and treatment choice.[1,3–6] Recent evidence suggests that adding a procedure to clarify values improves patient outcomes, but the impact emerges over time.[7,8] At present, many believe that supporting the process of values clarification is beneficial.[9]

Uncertainties confronted by healthcare providers and patients in the course of prostate cancer care indicate the need for improved measures to understand patient preferences.[3,4,10–13] This is particularly important because little is known about the optimal management strategies for prostate cancer and for advising patients regarding treatment choice.[3,4,10–13] The Institute of Medicine has defined patient-centered care as "providing care that is respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all decisions."[14] Patient-centered care that encompasses informed decision-making is a process of decision-making by patient and physician where the patient: 1) understands the risk or seriousness of the disease or condition to be prevented; 2) understands the preventive service, including risks, benefits, alternatives and uncertainties; 3) has evaluated his/her values regarding the potential benefits and harms associated with treatment; and 4) has engaged in decision-making at a level he/she desires and feels comfortable with.[4,12,15–17]

It is estimated that in year 2021, there will be 248,530 estimated new cases of prostate cancer, and 34,130 estimated deaths related to prostate cancer. Treatment decisions for prostate cancer patients are complicated. Men with early stage prostate cancer in particular face challenging treatment decisions since optimal treatment for prostate cancer remains unclear.[3,10–13,18–30] For patients with localized prostate cancer, treatment choices include active surveillance, watchful waiting, or aggressive, potentially curative therapies, such as radical prostatectomy (RP), including robotic-assisted laparoscopic prostatectomy (RALP), external-beam radiation therapy (EBRT), brachytherapy (BT) and proton therapy (PT), all with the potential for clinically significant side effects. Patient-centered care, a key component of high quality of care, involves application of scientific knowledge to patient care, tailored to each individual's unique characteristics, circumstances, needs and preferences.[3,4,7,10–13] In patient-centered prostate cancer care, concordance between patient preferences and treatment attributes may help optimize outcomes of care.[17,31–38]

Only few studies have assessed the patient treatment preferences, role desired by patients in decision-making, and the association of patient treatment preferences with outcomes among prostate cancer patients, as we do in this study. The objective of this randomized clinical trial is to study the association between preferences and outcomes of care among men with localized prostate cancer. We will also identify preferred attributes of alternative prostate cancer treatments (including active surveillance) that will aid in evaluating treatment options. Our study addresses the need for analyzing patients' values and their relationship to outcomes in order to determine if matching patients' values to attributes of treatment can improve quality of care and outcomes. Patient-centered care requires knowledge of how patients' preferences and reasoning affects choice of alternative therapeutic options. This study has direct relevance to patient-centered care as we will identify patient preferences, and how the concordance between preferences and attributes of treatment received affects the outcomes of prostate cancer treatment.

METHODS

Study Design and Conduct

The overall study methodology has been described previously.[39] Briefly, in this multi-centered RCT, the intervention was a web-based ACBC tool, Preferences for Prostate Cancer Care (PreProCare), for preferences assessment. The output from the tool was a list of the five attributes that the patient valued most. Patients completed the intervention either during the office visit or at home. All participants completed self-administered outcome assessments at baseline (prior to the intervention) and at 3, 6, 12 and 24-month follow-up. Participants were offered a $20 gift card at each assessment as a token of appreciation. Local institutional review boards approved the study. The study had a stakeholders advisory board and a data and safety monitoring committee.

Study Participants

Study Sites

The University of Pennsylvania (site 1) was the primary and coordinating site. Other study sites were the Corporal Michael J. Crescenz Veterans Administration Medical Center (site 2) and Fox Chase Cancer Center and Temple University Hospital (site 3). Based on sample size estimates, the total target accrual goal for our study was 720 participants.

Study Eligibility Criteria

The study eligibility criteria were: (1) newly diagnosed with localized prostate cancer (low risk: PSA ≤ 10 ng/ml, Gleason ≤ 6, and stage T1c–T2a; intermediate risk: PSA > 10–≤ 20 ng/ml, or Gleason 7, or stage T2b; and high risk: PSA > 20 ng/ml, or Gleason score 8–10, or stage T2c) group[10]; (2) treatment naive; (3) age ≥ 18 years; and (4) able to provide informed consent. The exclusion criteria were: (1) distant, metastatic or un-staged prostate cancer at diagnosis; (2) unable to communicate in English; and (3) already treated for prostate cancer.

Recruitment and Randomization

Recruitment involved following steps: 1) obtaining consent from the patient's urologist/physician for reviewing medical records; 2) determining eligibility via medical records; 3) screening to assess willingness to participate; and 4) obtaining informed consent and HIPAA permissions from participants. The study biostatistician (KM) created randomization sequences for each site using a pseudo-random number generator with random blocking varying in size from 2 to 6. The treatment assignments were placed in sealed, opaque envelopes. Research coordinators opened the envelope and notified participants of group assignment. Study investigators were masked to the treatment assignment. The study and consent form comply with HIPAA Standards for Privacy of Individually Identifiable Health Information. This study was approved by the local Institutional Review Board at all sites. The study was registered with Clinicaltrials.gov (NCT02032550).

Preferences for Prostate Cancer Care (PreProCare) Intervention

Participants in the intervention group completed the choice-based adaptive conjoint analysis instrument, PreProCare tool, to assess their individual preferences. The PreProCare was developed using Sawtooth Software. The details of development of this web-based instrument have been published previously.[40] As shown in Figures 1–4, briefly, in this three part tool, a brief introduction to the instrument is provided in part one. In the second part, the participants rank the attributes of various treatments ("not important" to "extremely important"). In the third part, choice scenarios consisting of combinations of attributes are presented based on the attributes ranking, with participants selecting the combination that they most prefer. At end of the task, a graph and a list of the five attributes most preferred by the participant is generated. The participant has the option to have a printout of the output to share with his provider. On average, this instrument required about 30 minutes to complete. Usual care group participants received care as usual that consisted of standard educational material about prostate cancer treatments.

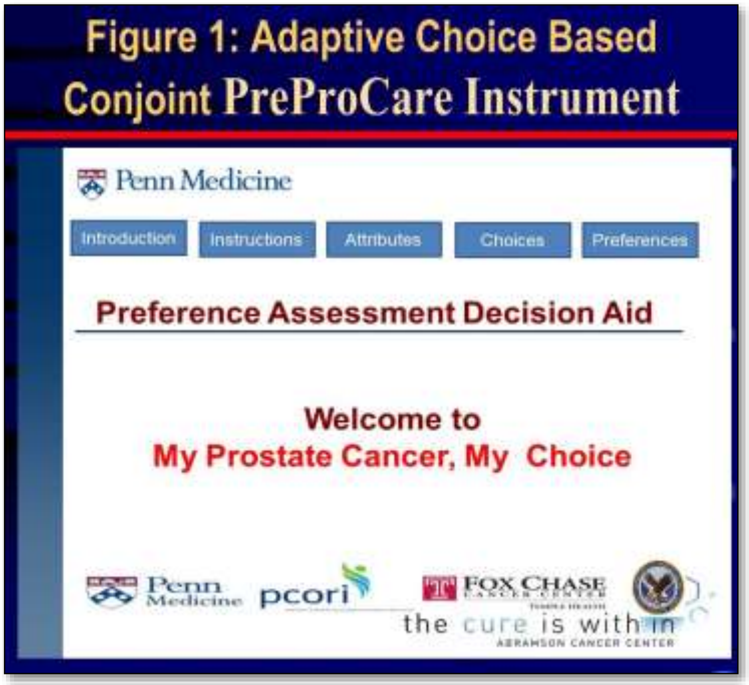


Figure 1: Adaptive Choice Based Conjoint PreProCare Instrument

Preference Assessment Decision Aid

Welcome to
My Prostate Cancer, My Choice

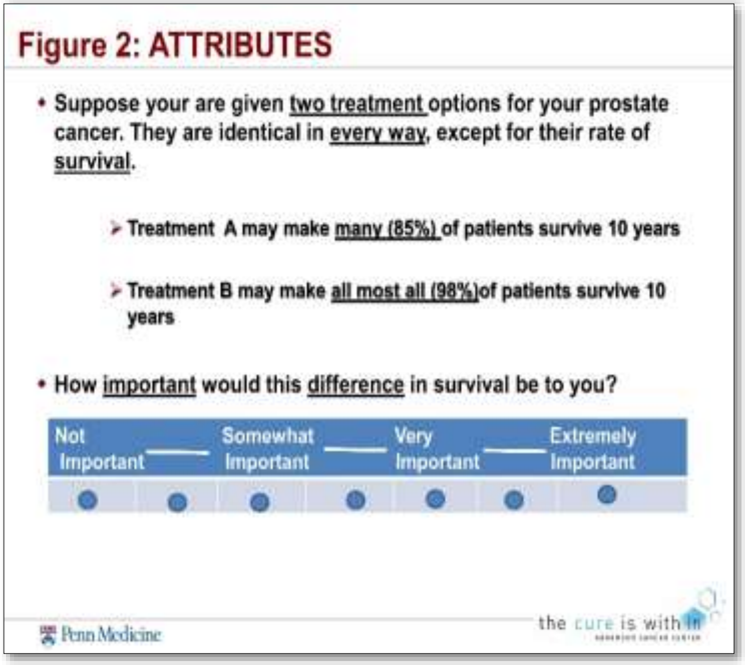


Figure 2: ATTRIBUTES

- Suppose your are given two treatment options for your prostate cancer. They are identical in every way, except for their rate of survival.
  - Treatment A may make many (85%) of patients survive 10 years
  - Treatment B may make all most all (98%)of patients survive 10 years
- How important would this difference in survival be to you?

Not Important		Somewhat Important		Very Important		Extremely Important
●	●	●	●	●	●	●

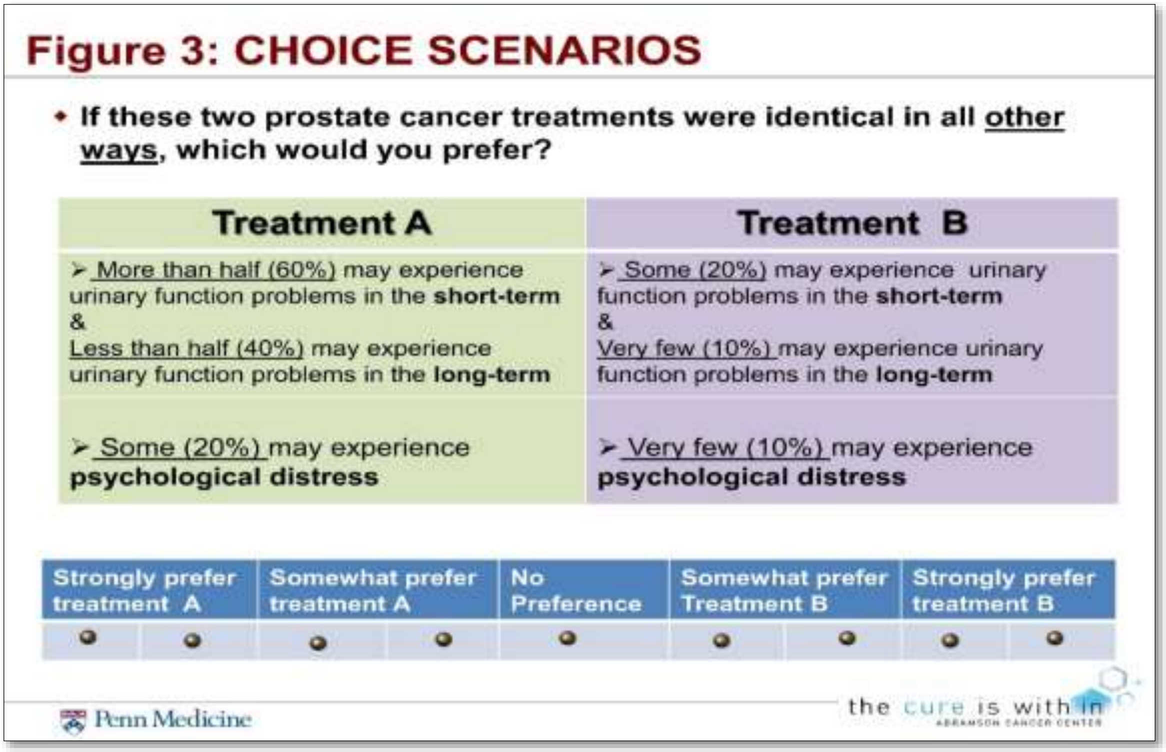


Figure 3: CHOICE SCENARIOS

- If these two prostate cancer treatments were identical in all other ways, which would you prefer?

Treatment A	Treatment B
➢ More than half (60%) may experience urinary function problems in the **short-term** & Less than half (40%) may experience urinary function problems in the **long-term**	➢ Some (20%) may experience urinary function problems in the **short-term** & Very few (10%) may experience urinary function problems in the **long-term**
➢ Some (20%) may experience **psychological distress**	➢ Very few (10%) may experience **psychological distress**

Strongly prefer treatment A	Somewhat prefer treatment A	No Preference	Somewhat prefer Treatment B	Strongly prefer treatment B
● ●	● ●	●	● ●	● ●

Penn Medicine

the cure is within
ABRAMSON CANCER CENTER

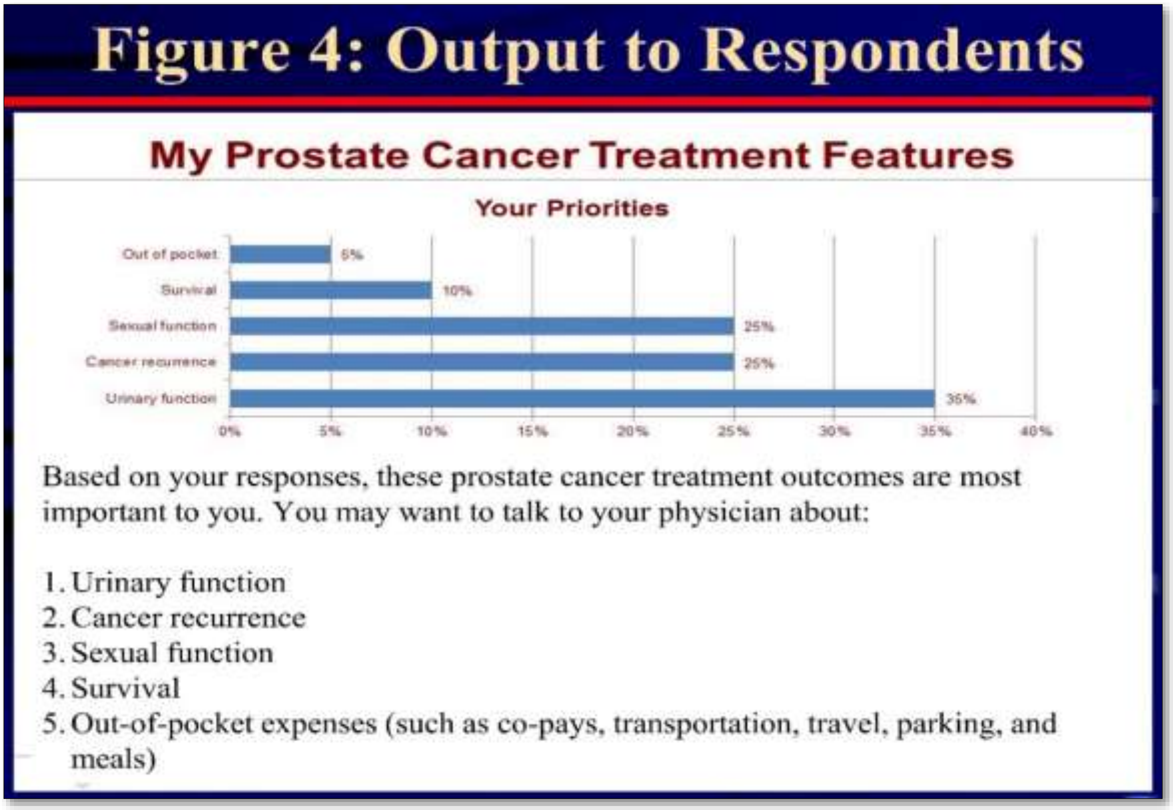


Figure 4: Output to Respondents

My Prostate Cancer Treatment Features

Based on your responses, these prostate cancer treatment outcomes are most important to you. You may want to talk to your physician about:

1. Urinary function
2. Cancer recurrence
3. Sexual function
4. Survival
5. Out-of-pocket expenses (such as co-pays, transportation, travel, parking, and meals)

Outcome assessments. The primary outcome was satisfaction with care, and secondary outcomes were satisfaction with decision, decision regret, and treatment choice. Satisfaction with care was measured at baseline and at follow-up. Satisfaction with decision and decision regret were measured at follow-up.

1. Satisfaction with care: Satisfaction with care is measured using Patient satisfaction questionnaire (PSQ-18).[41] In this questionnaire, 18 items are consolidated into seven subscales that assess satisfaction with medical care and six aspects of care. Scores on each subscale ranges from 1 to 5. Higher score demonstrates higher satisfaction. The PSQ-18 has good internal consistency (Cronbach's alpha, 0.86) and test-retest reliability (r, 0.92).
2. Satisfaction with decision: Satisfaction with decision is measured using the six-item Satisfaction With Decision (SWD) scale.[42] The total score ranges between 1 and 5, and higher scores indicate higher satisfaction with decision. The scale has excellent reliability (Cronbach's alpha, 0.88).
3. Decision regret: Decision regret is measured using the regret subscale of Memorial anxiety scale for prostate (MAX-PC).[43] The score on the five-item regret subscale ranges between 0 and 100. Higher score indicates higher decision regret. The scale has high internal consistency and validity.
4. Treatment choice: We obtained data on primary treatment, active surveillance, open radical prostatectomy, robot-assisted radical prostatectomy, and radiation therapy (intensity modulated, brachytherapy, or proton therapy) from self-report and from medical charts.

Covariates

Data on following potential confounding variables such as TNM stage of prostate cancer, Gleason score and Charlson comorbidity[44] were collected through medical chart review. We also obtained demographic data via standard self-report baseline data on patient age, income, race, ethnicity, education, health insurance, occupation, marital status, smoking status, height, weight and family history of prostate cancer.

Statistical Analysis

To begin with, we checked the data quality and carried out descriptive analyses of socio-demographic and outcome variables. We also assessed if there were differences in the characteristics of those who completed the study vs. those who were lost to follow-up. The outcomes were assessed at baseline, and at 3, 6, 12 and 24-month follow-up. We used an intent-to-treat approach to our analysis. To estimate longitudinal changes for satisfaction with care, satisfaction with decision, and decision regret, we used appropriate mixed effects models to adequately account for correlation among repeated measures. The models included fixed effects for group, time, and interaction of group-by-time, which allows for comparing the change in outcome over time between intervention group and usual care group. Both unadjusted and adjusted results were reported in terms of the Wald test, point estimates, and 95% confidence intervals. We also performed longitudinal comparison of mean scores for the two groups (intervention group vs. usual care group) for the subscales of satisfaction with care. For each of the six items of the satisfaction with decision (SWD), we compared the proportion satisfied between intervention group and usual care group.

Hierarchical Bayesian random effects regression analysis generated the utilities. Range of an attribute (differences between highest and lowest utility) divided by the sum of the ranges of all attributes yielded the percent of importance. We created three risk groups based on prostate cancer risk. For each risk group, treatment was compared between intervention group and usual care group. For the participants in the intervention group only, we developed values markers, assessed individual level utilities, and performed latent profile analysis based on the individual level utilities. Also, for the participants from intervention group, we modeled treatment as a function of importance of attribute, after adjusting for socio-demographic variables.

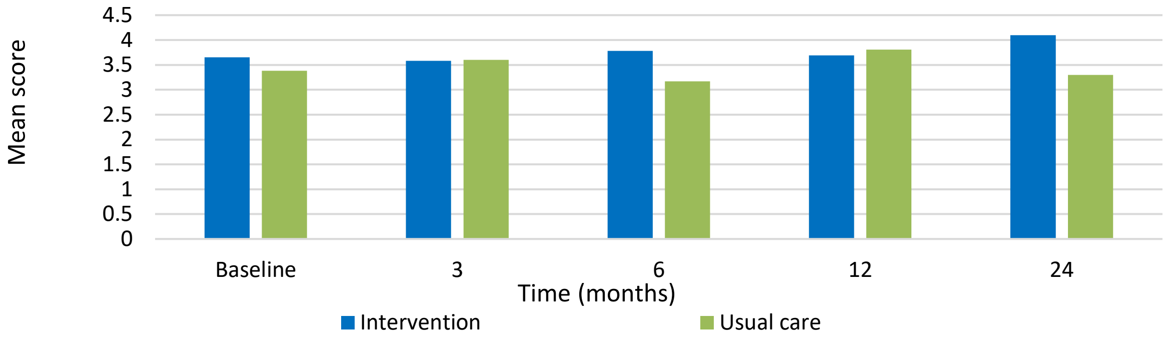
Results

We recruited a total of 743 localized prostate cancer patients between January 2014 and March 2015. Of these, 372 were randomized to the PreProCare intervention, and 371 were randomized to usual care. Overall retention was 74.2% at 24-months.

Satisfaction with Care: For the seven subscales of the PSQ-18, we estimated changes from baseline to 24-month follow-up and compared them between the intervention group and usual care group. Positive change indicates an improvement in satisfaction with care over time. All subscales of satisfaction with care demonstrated significant improvement for the intervention group, compared to the usual care group. For example, improvement in the mean score at 24-months from baseline for the general satisfaction subscale, was 0.44 (SE=0.06), or equal to 0.5 (SD) for the intervention group and was clinically and statistically significant.

Longitudinal Comparison of Mean Satisfaction with Care Scores: For the seven subscales of the PSQ-18, mean scores were compared between intervention and usual care group at baseline and at 3,6,12 and 24-month follow-up. In **Fig 5**, we present the longitudinal comparison of mean scores for the two groups for the general satisfaction subscale. It was observed that at the 24-month follow-up, the mean score on general satisfaction was higher for the intervention group, compared to the usual care group. We observed similar results for the other sub-scales.

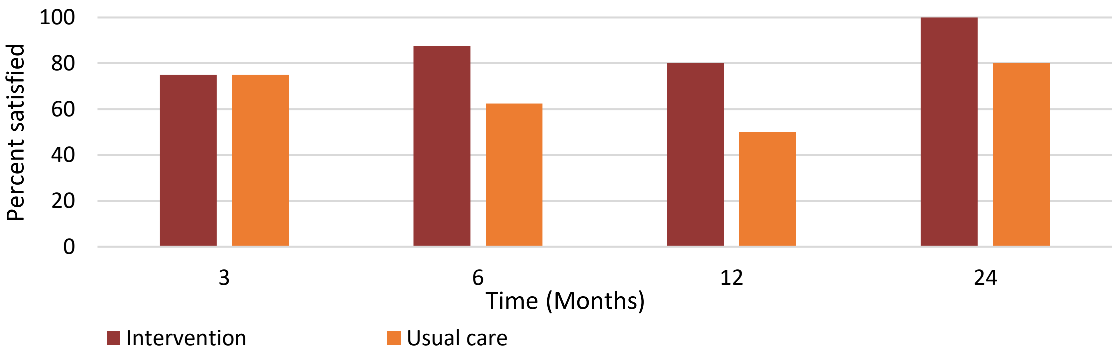
Fig 5: General satisfaction with care over time, by intervention status*



* Satisfaction with care scores range from1 to 5. Higher score indicates better satisfaction with care.

Satisfaction with Decision (SWD): Results of the log-gamma model of change in total score of SWD showed that satisfaction with decision had improved in the intervention group and also in the usual care group. However, the magnitude of improvement was larger and statistically significant for the intervention group. The estimate of change in mean score at 24-months from 3-month was -0.142 (SE=0.03) for the intervention group and -0.016 (SE= 0.03) for the usual care group. For each of the six items of the SWD, we compared the proportion satisfied between the intervention group and usual care group. As shown in **Fig 6**, for the item, “I am satisfied with my decision,” the proportion satisfied at 3 months was comparable between the two groups. However, at 6, 12 and 24 months a higher proportion of participants from the intervention group reported satisfaction with decision, compared to usual care group (all p values < 0.05). A similar pattern was observed for the other items of the SWD.

Fig 6: Satisfaction with decision over time, by intervention status*



* I am satisfied with my decision

Decision Regret: Results of the mixed effects model for change in regret score indicated that regret had reduced in both groups (intervention group and usual care group). However, the reduction was larger for the intervention group. The change in estimated mean score at 24 months from 3 months was -2.52 (SE=1.27) for the intervention group and 0.61 (SE=1.23) for the usual care group. Negative change is indicative of decline in regret.

Treatment Choice by Prostate Cancer Risk Category: In the low-risk group participants, a higher proportion from the intervention group were on active surveillance compared to the usual care (66% vs. 54%; p=0.24). Among participants from the intermediate-risk group, the proportion on active surveillance was comparable (12% vs. 13%, p=0.76). In participants from the high-risk category, 93% of the intervention group received active treatment, compared to 95% from the usual care group (p=0.72).

Table 1: Association between Utilities and Type of Treatment

Attributes	**Treatment Type**		
	Radiation	Surgery	Active Surveillance
	Odds Ratio (95% CI)	Odds Ratio (95% CI)	Odds Ratio (95% CI)
Survival	1.20 (1.06, 1.50)	1.11 (1.06, 1.28)	0.67 (0.48, 0.94)
Cancer recurrence or progression	0.90 (0.70, 1.16)	0.99 (0.86, 1.15)	1.02 (0.75, 1.38)
Change in urinary function	1.07 (0.85, 1.36)	0.98 (0.85, 1.15)	1.11 (0.77, 1.60)
Change in bowel function	1.18 (1.03, 1.51)	0.95 (0.82, 1.11)	1.09 (0.75, 1.58)
Change in sexual function	0.95 (0.75, 1.19)	0.93 (0.80, 0.99)	1.46 (1.04, 2.06)
Psychological distress	1.22 (1.09, 1.56)	0.90 (0.78, 0.99)	0.88 (0.63, 1.23)
Side effects	0.71 (0.54, 0.95)	0.99 (0.85, 1.16)	0.99 (0.67, 1.47)
Treatment duration	0.87 (0.66, 1.16)	0.98 (0.83, 1.15)	1.11 (0.78, 1.58)
Need for cutting	1.05 (0.84, 1.31)	0.88 (0.77, 0.99)	1.54 (1.06, 2.24)
Radiation or seed implants	0.86 (0.67, 1.10)	1.01 (0.88, 1.17)	1.10 (0.80, 1.51)
Recovery time	0.98 (0.76, 1.27)	0.98 (0.85, 1.15)	1.43 (1.02, 1.99)
Cancer control	0.90 (0.71, 1.14)	1.16 (1.00, 1.34)	1.10 (0.81, 1.51)
Out-of-pocket expenses during treatment year	1.12 (0.89, 1.38)	0.98 (0.86, 1.13)	1.03 (0.77, 1.37)
Caregiver burden	0.84 (0.64, 1.11)	1.03 (0.88, 1.21)	1.09 (0.77, 1.55)

Association between Attribute and Treatment Type: Results of the logistic regressions among 371 prostate cancer patients from the intervention group are presented in Table 1. We observed that survival was associated with higher odds of radiation treatment (OR=1.20, CI=1.06, 1.50). Survival was also associated with higher odds of surgery (OR=1.11, CI=1.06, 1.28). On the other hand, survival was associated with lower odds of being on active surveillance (OR=0.67, CI=0.48, 0.94). Sexual function was associated with higher odds of being on active surveillance (OR=1.46, CI=1.04, 2.06). Fear of cutting was associated with lower odds of surgery (OR=0.88, CI=0.77, 0.99). Finally, fear of cutting was also associated with higher odds of being on active surveillance (OR=1.54, CI=1.04, 2.24).

Discussion

Preference assessment is the cornerstone of patient-centered care and has been universally accepted as a method to improve the quality of patient care. However, evidence that preference assessment itself improves patient-centered outcomes and treatment choice is lacking. In our novel patient-centered RCT of more than 700 patients with localized prostate cancer, we observed that our preference assessment intervention, the PreProCare tool, was associated with improved satisfaction with care, increased decision satisfaction, lower regrets, and treatment choices were more consistent with the estimated risk level of prostate cancer diagnosis, clearly stated values/preferences for the outcomes that might be experienced. Value markers or utility levels of treatment such as survival, recovery time and sexual function were associated with active surveillance among low-risk prostate cancer patients. Our study is the first to analyze the association between conjoint analysis value assessment intervention and outcomes and determined (1) the comparative effectiveness of preference assessment intervention on treatment choice outcomes vs. usual care; (2) the importance of an assessment of value markers in patient-centered care; and (3) association between value markers (utility levels) and treatment choice. Our study demonstrated that in localized prostate cancer, helping patients identify their own preferences using a structured, standardized computer-based preference assessment tool might be a mechanism for enhancing patient-centered decision-making and outcomes.

Conclusions

Preference assessment intervention can help prostate cancer patients reveal their preferences, leading to better alignment with treatment decision. Future research should identify strategies to ensure diagnosis and treatment options are communicated to patients accurately.

Acknowledgement

Supported by the Patient-centered Outcomes Research Institute—CE-12-11-4973.

Ravishankar Jayadevappa

References

1. Stacey D, Légaré F, Lewis K, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane Database of Systematic Reviews.* 2017;4 (Art. No.: CD001431).

2. Barnato AE, Llewellyn-Thomas HA, Peters EM, Siminoff L, Collins ED, Barry MJ. Communication and Decision Making in Cancer Care: Setting Research Priorities for Decision Support/Patients' Decision Aids. *Medical Decision Making.* 2007;27(5):626–634.

3. Gwyn R, et al. When is a shared decision not [quite] a shared decision? Negotiating preferences in a general practice encounter. *Soc Sci Med.* 1999;49 437–447.

4. Institute of Medicine. Crossing the Quality Chasm: A new health system for the 21st century. Washington (DC): NAP;2001.

5. IOM (Institute of Medicine). *Patient-Centered Cancer Treatment Planning: Improving the Quality of Oncology Care: Workshop Summary.* Washington, DC: The National Academies Press;2011.

6. Pauker SG. Medical Decision Making: How Patients Choose. *Medi Decis Making.* 2010;30(suppl 1):8s–11s.

7. Feldman-Stewart D, et al. A conceptual framework for patient–professional communication: an application to the cancer context. *Psycho-oncology.* 2005;14:801–809.

8. Feldman-Stewart D, Tong C, Siemens R, et al. The impact of explicit values clarification exercises in a patient decision aid emerges after the decision is actually made: evidence from a randomized controlled trial. *Med Decis Making.* 2012;32(4):616–626.

9. Elwyn G, Stiel M, Durand MA, Boivin J. The design of patient decision support interventions: addressing the theory-practice gap. *J Eval Clin Pract.* 2011;17(4):565–574.

10. D'Amico AV, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. *JAMA.* 1998;280:969–974.

11. Eraker SA, Sox HC Jr. Assessment of patients' preferences for the therapeutic outcomes. *Med Decis Making.* 1981;1 (1):29–39.

12. Stewart M ea. Patient-centered medicine: transforming the clinical methods. In. Vol London: SAGE; 1995.

13. Stewart ST ea. Utilities for prostate cancer health states in men aged 60 and older. *Medical Care.* 2005;43 (4):347–355.

14. Institute of Medicine. *Policy issues in the development of personalized medicine in oncology: Workshop summary.* Washington, DC: Institute of Medicine;2010.

15. Benbassat J, et al. Patient's preference for participation in clinical decision-making: a review of published surveys. *Behav Med.* 1998;24:81–88

16. Guadagnoli E, et al. Patient participation in decision making. *Soc Sci Med.* 1998;47:329–339.

17. Jayadevappa R, Chhatre S. Patient Centered Care—A Conceptual Model and Review of the State of the Art. *The Open Health Services and Policy Journal.* 2011;4:15–25.

18. Sommers BD, et al. Decision analysis using individual patient preferences to determine optimal treatment for localized prostate cancer. *Cancer.* 2007;110: 2210–7:2210–2217.

19. Merrick GS, et al. Long-term urinary quality of life after permanent prostate brachytherapy. *Int J Radiation Oncology Biol Phys.* 2003;56(2):454–461.

20. Albertsen PC. 20-year outcomes following conservative management of clinically localized prostate cancer. *JAMA.* 2005;293 (17):2095–2101.

21. Caffor O, et al. Prospective evaluation of quality of life after interstitial brachytherapy for localized prostate cancer. *Int J Radiation Oncology Biol Phys.* 2006;59(2):1532–1538.

22. Sanda MG, et al. Quality of life and satisfaction with care outcome among prostate cancer survivors. *NEJM.* 2008;358 1250–1261.

23. Boehmer U, Babayan RK. Facing erectile dysfunction due to prostate cancer treatment: perspectives of men and their partners. *Cancer Investigation.* 2004;22(6):840–848.

24. Clark JA, et al. Living with treatment decisions: regrets and quality of life among men treated for metastatic prostate cancer. *Journal of Clinical Oncology.* 2001;19 (1):72–80.

25. Frosch DL, et al. Internet patient decision support-a randomized controlled trial comparing alternative approaches for men considering prostate cancer screening *Arch intern Med.* 2008;168(4):363–369.

26. Penson DF, et al. Health related quality of life in men with prostate cancer. *The Journal of Urology.* 2003;169.

27. Davison BJ, et al. Information and decision making preferences of men with prostate cancer. *Oncol Nurs Forum.* 1995;22:1404–1408.

28. Jayadevappa R, Bloom BS, Chhatre S, Fomebrstein KM, Wein AJ, Malkowicz SB. Health related quality of life and direct medical care cost of newly diagnosed younger men with prostate cancer. *The Journal of Urology.* 2005;174:1059–1064.

29. Jayadevappa R, Chhatre S, Bloom BS, Whittington R, Wein A, Malkowicz SB. Health related quality of life and satisfaction with care among older men treated with radical prostatectomy or external beam radiation therapy. *British Journal of Urology International.* 2006; 97 955–962.

30. Wilt TJ, Ullman KE, Linskens EJ, et al. Therapies for Clinically Localized Prostate Cancer: A Comparative Effectiveness Review. *the Journal of Urology.* 2021;205 967–976.

31. Feldman-Stewart D, Brundage MD. A conceptual framework for patient–provider communication: a tool in the PRO research tool box. *Quality of Life Research.* 2009;18:109–114.

32. Zeliadt SB, et al. Preliminary treatment considerations among men with newly diagnosed prostate cancer. *American Journal of Managed Care.* 2010;16(5):e121–130.

33. Kramer KM, et al. Patient Preferences in prostate cancer: a clinician's guide to understanding health utilities. *2005.* 2005;4(1):15–23.

34. Bridges JF, al e. Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value in Health.* 2011(in press).

35. Lin GA, et al. Patient Decision Aids for prostate cancer treatment A Systematic Review of the Literature. *CA Cancer J Clin.* 2009;59:379–390.

36. Feldman-Stewart D, et al. A decision aid for men with early stage prostate cancer: theoretical basis and a test by surrogate patients. *Health Expectations.* 2001;4:221–234.

37. Bond C. *Concordance A Partnership in Medicine Taking.* London: Pharmaceutical Press; 2004.

38. Elstein AS, Chapman GB, et al. Agreement between prostate cancer patients and their clinicians about utilities and attribut importance. *Health Expectations.* 2004;7:115–125.

39. Jayadevappa R, Chhatre S, Gallo JJ, et al. Patient-Centered Preference Assessment to Improve Satisfaction With Care Among Patients With Localized Prostate Cancer: A Randomized Controlled Trial. *Journal of Clinical Oncology.* 2019;37:1–13.

40. Jayadevappa R, Chhatre S, Gallo JJ, et al. Patient-Centered Approach to Develop the Patient’s Preferences for Prostate Cancer Care (PreProCare) Tool. *MDM P&P.* 2019:1–13.

41. Ware JE ea. *Development and validation of scales to measure patient satisfaction with medical care services.* Springfield, VA,: National Technical Information Service, 1976.;1976.

42. HOLMES-ROVNER M, Kroll J, Schmitt N, et al. Patient Satisfaction with Health Care Decisions: The Satisfaction with Decision Scale. *Medical decision making.* 1996;16:56–64.

43. Roth A, Nelson CJ, et al. Assessing Anxiety in Men With Prostate Cancer: Further Data on the Reliability and Validity of the Memorial Anxiety Scale for Prostate Cancer (MAX–PC). *Psychosomatics.* 2006;47:340–347.

44. Singh R, O’Brien TS. Comorbidity assessment in localized prostate cancer: a review of currently available techniques. *European Urology.* 2004;46:28–41.

PRICE FIXING: CAN YOU CONSTRUCT AN EXPERIMENTAL DESIGN TO MAKE PRICE-SENSITIVE CONSUMERS APPEAR PRICE SEEKING?

JAKE LEE
RED ANALYTICS, INC

ABSTRACT

Experimental design is a necessary component of choice experiments. It is the design that determines which combinations of features make up the alternatives and which collection of alternatives make up each task. An efficient experimental design balances the need of the analyst to convert the consumer choice process into a mathematical form and the desire to keep the consumer survey time to a minimally acceptable length.

Analysts observe weird model results on occasion and are asked to guess what they think is going on. It is not always easy to determine if consumers are just weird or if there was a misstep in the analysis process. A defective experimental design should be considered as a possible culprit when unintuitive results are spotted.

In this paper we will review an experimental design that completely ruined the inference of the study. We will go in depth on what the issues were and provide some suggestions to avoid disaster in the future. If you would like to follow along, the experimental design file can be found at https://redanalytics.net/dumpsterfire. To keep the project confidential there are no attribute or level labels in the file, but we will do our best to describe the necessary components.

In the last section we will introduce the error cliff as a framework to help avoid design mistakes that result from clients or bosses pushing to include too many attributes or provide too few tasks.

BACKGROUND

A friend was conducting a choice experiment as part of a consulting engagement. When he saw the perpetually increasing price-revenue curves, he reached out to our analytics group to see if we had any tips to help their analyst fix the problem.

Initial modeling showed around one-third of the subjects had a positive price parameter. The model was suggesting that a large number of people preferred higher prices. The range of prices tested was very wide and there was no rational explanation as to why anyone would prefer a higher price.

The client was paying between six and seven figures for product guidance, and the suggestion that they can profitably raise price in perpetuity was laughable on its face. The consultants knew that comments like “Well, that is what the model says” are on the fast track to getting thrown out of the office and never rehired for more consulting.

Project Specifics

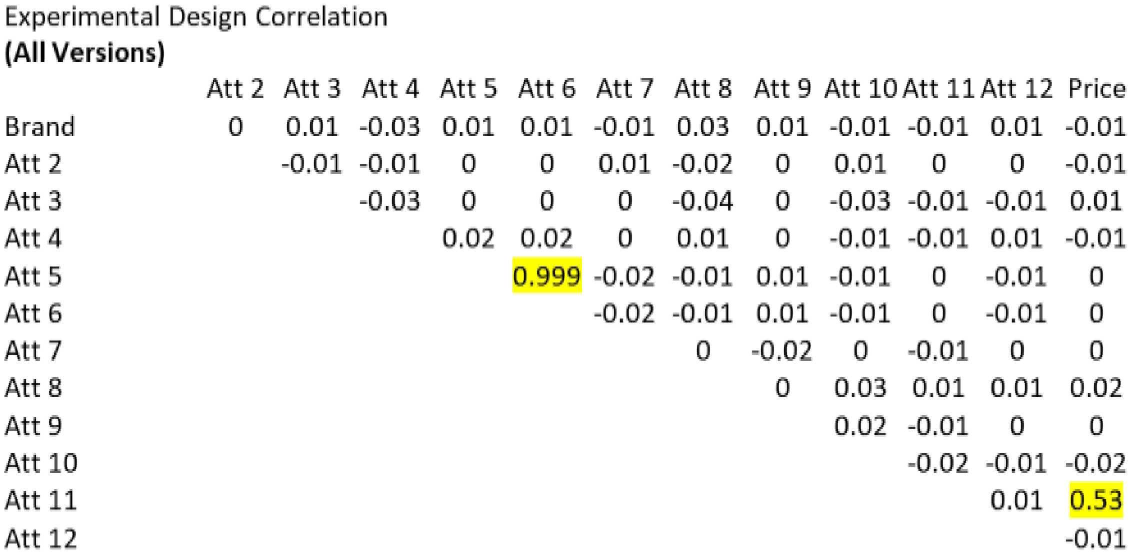
The category studied in this case is confidential. There were a total of 13 attributes. Brand had seven levels and price had five. Two of the attributes had eight levels, one had three, another eight attributes were binary.

There were prohibitions between attributes five and six and between attribute eleven and price. The prohibitions were completely unnecessary, but it is understandable that someone would suggest including them to make the tasks more "realistic."

The survey had 300 versions of the choice tasks. Each respondent was randomly assigned to a version with ten tasks. Each task has three alternatives and a dual response none. They collected responses from over 2,400 people, which is more than enough for the model.

The decision criteria used for the number of tasks, alternatives and number of respondents used by the consultants is unclear.

Figure 1



Experimental Design Correlation
(All Versions)

	Att 2	Att 3	Att 4	Att 5	Att 6	Att 7	Att 8	Att 9	Att 10	Att 11	Att 12	Price
Brand	0	0.01	-0.03	0.01	0.01	-0.01	0.03	0.01	-0.01	-0.01	0.01	-0.01
Att 2		-0.01	-0.01	0	0	0.01	-0.02	0	0.01	0	0	-0.01
Att 3			-0.03	0	0	0	-0.04	0	-0.03	-0.01	-0.01	0.01
Att 4				0.02	0.02	0	0.01	0	-0.01	-0.01	0.01	-0.01
Att 5					0.999	-0.02	-0.01	0.01	-0.01	0	-0.01	0
Att 6						-0.02	-0.01	0.01	-0.01	0	-0.01	0
Att 7							0	-0.02	0	-0.01	0	0
Att 8								0	0.03	0.01	0.01	0.02
Att 9									0.02	-0.01	0	0
Att 10										-0.02	-0.01	-0.02
Att 11											0.01	0.53
Att 12												-0.01

On deeper inspection, the levels were perfectly balanced. That is, within an attribute each level appeared the same number of times across the design file. Figure 1 shows the design file was orthogonal (uncorrelated) with a few exceptions (highlighted) produced by the prohibitions. Level overlap was fully minimal. Except for the binary attributes, a single task never had the same level show up more than once.

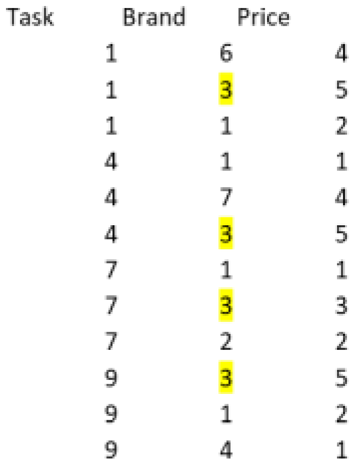
Diagnosing the Issues

The severe design issues were hidden, but only slightly, below the surface. Inspection or diagnostics aggregated across views obscured the deficits within each block. The file did not appear to have any issues upon the initial inspection. But the resulting inference was unbelievable, so a deeper look was needed.

The first clue came when looking at the brand part-worths from the initial model results. Many of the values were in the teens. And in logit space, it is surprising (not impossible) to see values outside of +/-4. The question arose: what exactly did the first subject respond to that led to such a high part-worth value for brand 1?

The respondent had chosen brand 1 every time it appeared in a task—a total of four times. Was there anything special about those four tasks where brand 1 was chosen that could lead to the extreme part-worths? Yes. Brand 3 appeared with brand 1 every time. Most versions of the design had the same pattern with two of the seven brands co-occurring in every task. See Figure 2 for the tasks in version 1 that included brand 1 as an option.

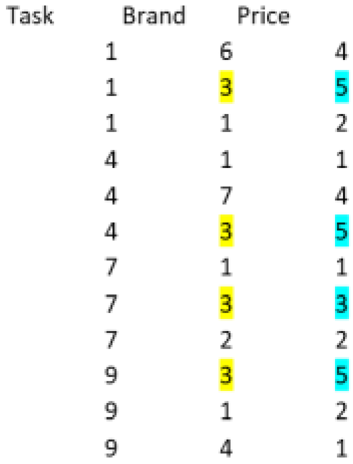
Figure 2



Task	Brand	Price
1	6	4
1	3	5
1	1	2
4	1	1
4	7	4
4	3	5
7	1	1
7	3	3
7	2	2
9	3	5
9	1	2
9	4	1

But wait—it gets worse. Another pattern that is even harder to detect was also in version 1 of the design. Brand 3 was always assigned a higher price than brand 1. Every time. See Figure 3 for the same set of tasks and the higher price for brand 3 highlighted. Use the link to the full design in the abstract if you want to see how this pattern emerges throughout the entire design file.

Figure 3



Task	Brand	Price
1	6	4
1	3	5
1	1	2
4	1	1
4	7	4
4	3	5
7	1	1
7	3	3
7	2	2
9	3	5
9	1	2
9	4	1

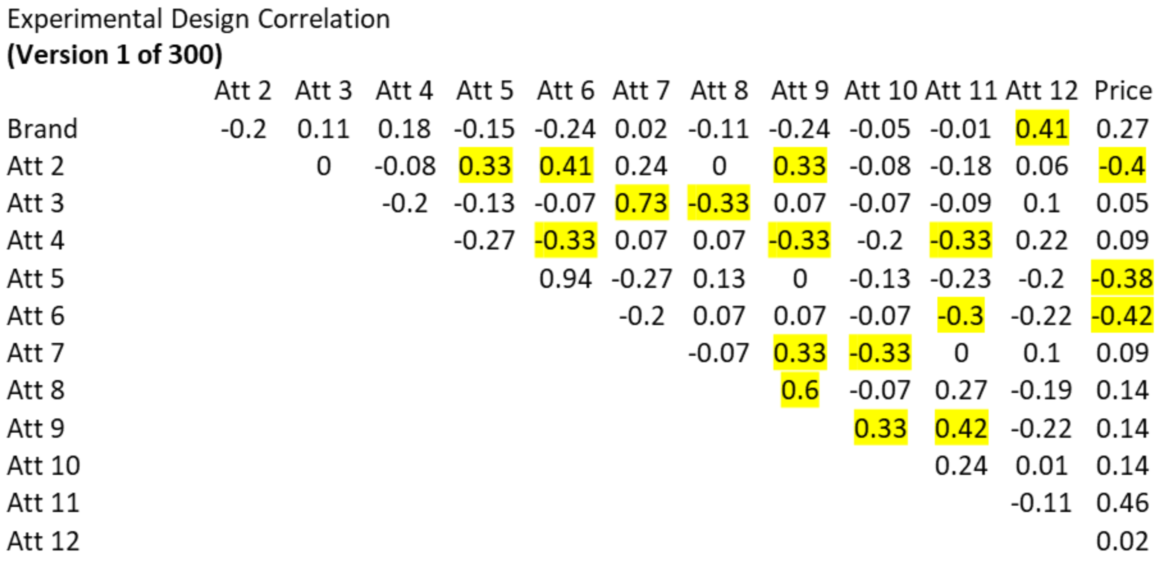
In version 1, brands 1 and 3 were confounded with price. There was no way for the statistical model to know if a respondent picking brand 1 in each task was choosing it because of the brand or the lower prices. Also, the statistical model is not smart enough to know that a person choosing brand 3 every time probably likes the brand and not the higher prices.

Roughly 60% of the tasks had a fully co-occurring, price-dominated, brand pair. Since the hierarchical Bayes (HB) model evaluates the likelihood function at the individual level, this is not something that simply averages out across design versions. This confounding drove both the apparent lack of price sensitivity and the extreme brand coefficients.

An experimental design procedure that produces this type of pattern cannot be considered random, despite the analyst's insistence that he used the software options for a "random" design.

The design was not as orthogonal as we first suspected. When you look at each version of the design independently, major dependencies begin to emerge. Figure 4 shows extreme correlations between the attributes when looking at version 1 of the design. Since each respondent only sees a single version of the design, it makes more sense to construct/diagnose the design one version at a time.

Figure 4



Experimental Design Correlation
(Version 1 of 300)

	Att 2	Att 3	Att 4	Att 5	Att 6	Att 7	Att 8	Att 9	Att 10	Att 11	Att 12	Price
Brand	-0.2	0.11	0.18	-0.15	-0.24	0.02	-0.11	-0.24	-0.05	-0.01	0.41	0.27
Att 2		0	-0.08	0.33	0.41	0.24	0	0.33	-0.08	-0.18	0.06	-0.4
Att 3			-0.2	-0.13	-0.07	0.73	-0.33	0.07	-0.07	-0.09	0.1	0.05
Att 4				-0.27	-0.33	0.07	0.07	-0.33	-0.2	-0.33	0.22	0.09
Att 5					0.94	-0.27	0.13	0	-0.13	-0.23	-0.2	-0.38
Att 6						-0.2	0.07	0.07	-0.07	-0.3	-0.22	-0.42
Att 7							-0.07	0.33	-0.33	0	0.1	0.09
Att 8								0.6	-0.07	0.27	-0.19	0.14
Att 9									0.33	0.42	-0.22	0.14
Att 10										0.24	0.01	0.14
Att 11											-0.11	0.46
Att 12												0.02

The design procedure masked serious design flaws within each version by averaging out the problems across 300 versions.

Experimental Design Concepts

The central idea in experimental design is efficiency. For a given number of tasks (or survey time) you want to minimize the error in the model that is due to the design. Different combinations of features/profiles will lead to different levels of error in the final model.

Kuhfeld (1994) observed that efficient designs for the general linear model (GLM) are both orthogonal and balanced. Orthogonal means the profiles are uncorrelated and balanced means each level within an attribute appears with equal frequency. In a multi-profile setting, which is typical in most choice experiments, orthogonal also means minimum overlap. That is, each level will appear as few times as possible per task.

In addition to orthogonality and balance, utility balance (not to be confused with level balance) has been suggested to improve efficiency (Huber 1996). Also, a moderate amount of overlap might improve study outcomes (Chrzan 2010).

In choice experiments, the multinomial logit model (MNL) differs from the standard GLM. For the MNL the design efficiency depends on the model parameters. Which is kind of odd because, if we knew the model parameters, we would not need to do the experiment in the first place.

A major criticism of characteristic-based designs is a lack of priority among the characteristics. There are measurements for orthogonality, overlap and balance. You can compare two designs and see how they differ on each of these three dimensions separately. But how do you improve a design in all three dimensions? Sometimes they disagree—an improvement in one dimension can easily be a detriment in another. When they do disagree, which one takes priority over the others?

Alternatively, optimal designs use a single measure of efficiency to iterate through candidate changes and improve the existing design. D-efficiency is the most common metric, but other measures are possible for special cases. D-efficiency seeks to minimize the "size" of the model's covariance matrix. The determinant of the covariance matrix is what the process seeks to minimize as it swaps profiles or levels depending on the algorithm used.

The largest hurdle for optimal designs is the need to specify a matrix of parameters to optimize around. In most cases an expert's educated guess on likely parameters will do.

USER ERROR

The analyst made some understandable mistakes in generating the design. Attributes 5 and 6 should have been combined to make a three-level attribute instead of two constrained binary attributes. The other prohibitions were also unnecessary. Combinations that were unlikely to be chosen by the brand decision makers were restricted in the design generation. In most cases strong priors on the parameters are preferred over strict prohibitions in the design space.

Despite these errors, these mistakes seem unlikely candidates for causing the price-dominated brand pairs.

They also needed more tasks. A common, yet not universally known, heuristic states that each level should appear at least six times. The eight level attributes appeared only 3.75 times each. 10 tasks * 3 alternatives / 8 levels = 3.75 appearances/level. The six appearances per level heuristic is not foolproof, but a good place to start.

If the user intended on a maximum likelihood model for the parameter estimation, using a large number of respondents and few tasks would be appropriate. But for the heterogeneous HB model, more tasks are usually needed when the design space gets large.

Bad Design Consequences

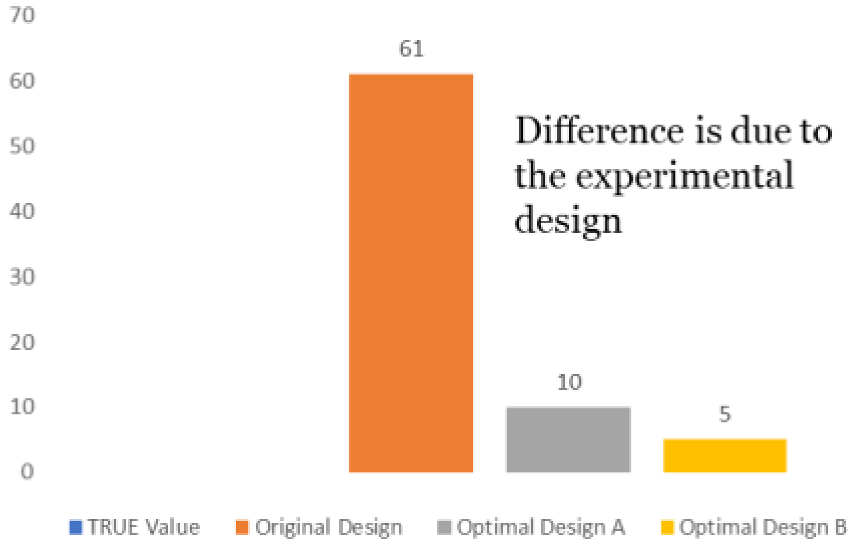
We set up a simulation experiment to explore how bad the actual design was relative to an optimal design with the same setup specifications. We wanted to see if more sims would appear price seeking depending on the design in front of them.

The strategy was to start with the same parameters from the best model results from the original data set. Then we shrunk the variance on the price parameters so that values for each sim were negative. We generated 2,400 sims and verified all had negative utility for higher prices.

The experiment had each of the sims complete multiple-choice tasks. The choices were based on highest utility plus Gumbel error. We treated them with three different designs. First, the original design. Second, a D-efficient design. Third, a D-efficient design with an extra alternative and two extra tasks. More on this third design when we cover the error cliff in a moment.

The key statistic we were concerned about was how many sims that have a confirmed negative price coefficient, ended up with a false positive price coefficient. Figure 5 shows that the "random" design led to 6 times as many false positives in the price parameter compared to the optimal design with the same specifications. Adding a few tasks, the model further reduced the number of false positives in half.

Figure 5



The same sims were used in all three treatments. So the differences in the amount of apparent price-seeking subjects is due to the experimental design strategy.

RECOMMENDATIONS

The first recommendation is more robust testing procedures. Estimating a model with fake responses to see if it "runs" will catch only the most egregious design mistakes. This strategy is not sufficient to spot price-dominated brand pairs. HB almost always works even when the data is very poorly conditioned. All tests should be done at the version level and not across all versions.

The author is not aware of a robust set of design checking procedures. A comprehensive and standardized protocol would be appreciated in the future.

The second recommendation is to consider upgrading to optimal experimental design procedures. Optimal designs should never have price-dominated brand pairs, because doing so is inefficient. Even with very loose and reasonable priors on the parameter set, the algorithm will replace price-dominated brand pairs with a set of product profiles more suited to minimize uncertainty in the model parameters. Even though the optimal design procedures should avoid many design problems, robust testing is still recommended.

The third recommendation is use simulation to discover the best trade-offs in design efficiency. This is illustrated in the error cliff.

THE ERROR CLIFF

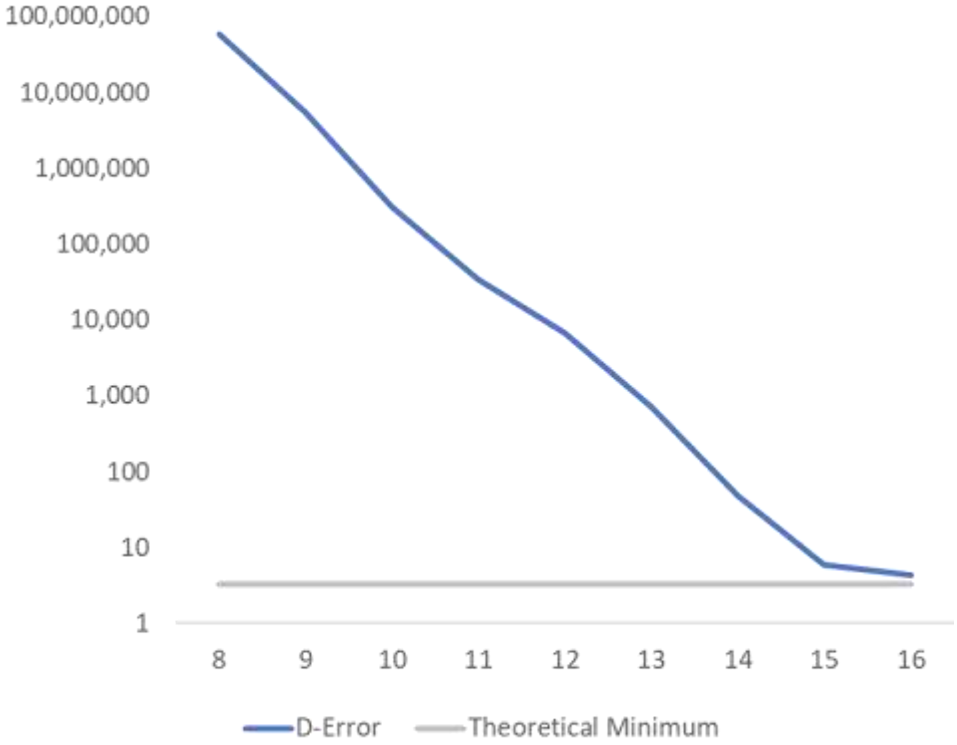
The error cliff was invented when a competing analytics group told a client they could have up to 120 parameters in their choice experiment, and there was not a maximum number of levels per attribute. Obviously, they are using some other set of statistical foundations to spec out experiments. Maybe a comedy book?

Humor aside, it does beg the question: how do you know when you have too many attributes or levels or when do you have too few tasks? In this specific case we are shy of the six appearances per level heuristic. But can we quantify that and determine a more precise cut-off?

We set up an experimental design simulation, using all of the inputs from the case study. We created optimal experimental designs while increasing the number of tasks from eight to sixteen. For each optimal design we recorded and charted the D-Error to visually inspect the trade-off between tasks and model error due to the design. See Figure 6.

The steep decent bends creating an elbow around 15 tasks. Which coincidentally is right where you would be if you used the six appearances per level heuristic. The chart is shown on log scale so that you can actually see the lines.

Figure 6



The same analysis showed that a dozen 4-up tasks would also reach safe efficient ground, which is why we included that as an alternative simulation result from Figure 5.

This technique can be used to understand when you are on solid footing whether you are wondering if you can get away with less tasks, or if you can add even more brands into the design space.

Conclusion

If a poor experimental design can change the sign on price parameters, it can ruin the inference and any recommendations of the study. Optimal experimental design procedures should be less likely to produce a bad study, however, robust design checking is suggested regardless of the design procedure chosen.

Optimal designs have the advantage of generating an error cliff to explore the trade-offs between levels and tasks.

Jake Lee

Bibliography

Chrzan, Zepp, White (2010) The Success of Choice-Based Conjoint Designs Among Respondents Making Lexicographic Choices, *Sawtooth Software Conference*

Kuhfeld, W.F., Tobias, R.D., Garratt M. (1994) Efficient Experimental Design with Marketing Research Applications, *Journal of Marketing Research*

Huber, Joel, Klaus Zwerina (1996) The Importance of Utility Balance in Efficient Choice Designs, *Journal of Marketing Research*

Street, Deborah J. and Leonie Burgess (2007) The Construction of Optimal Stated Choice Experiments: Theory and Methods, *Hoboken: Wiley*

SHOULD SHAPLEY-ADJUSTED REGRESSION COEFFICIENTS BE USED IN MAKING PREDICTIONS?

JACK HORNE[1]
JACK HORNE CONSULTING

INTRODUCTION

Customer satisfaction is one of the most widely measured concepts in market research. Efforts to measure and track customer satisfaction often consume most of the market research budget for many firms. Just as prevalent is the need to understand what the key drivers of consumer satisfaction are.

Say, for example that we probe for satisfaction on three elements: overall technical support (v_1), product functionality (v_2), and the instruction manual (v_3). We collect data for each of these measures, including overall satisfaction, on 10-point scales. To determine how each of these elements affects overall satisfaction, we run linear regression analysis, using overall satisfaction as the dependent variable, and the three service/product elements as predictors. The analysis results in regression coefficients for each element, which can be rescaled to find the relative importance of each element in explaining overall satisfaction.

So far, so good. We have a model that seeks to *predict* overall satisfaction from satisfaction with specific elements of the service or product, and that tells us the *relative importance* of each of those elements in determining overall satisfaction. We can use both ideas to focus our efforts in improving customer satisfaction. This is a straightforward approach and seems practical to answering our questions about what is affecting satisfaction. Unfortunately though, it has a significant likelihood of failing (Stratmann et al., 1994).

Linear regression in survey-based customer satisfaction research has several shortcomings in actual practice. Regression models optimize *predictions*, not relative importance of the contributing variables. If two highly correlated variables both explain variance in some other variable in similar ways, then once the first has entered the model and contributed its explanatory power, there isn't much to be added by inserting the second variable. This often causes the regression coefficient for the first variable to be large, and the coefficient for the second variable to be near zero, or vice versa. From the point of view of relative importance, such a result would lead us to conclude that the first variable is highly important in explaining the dependent, and the second is not, even though the two variables have similar correlations with the dependent variable.

At times, we may even find that the regression coefficient for some variables has the opposite sign compared to the direction of the pairwise correlation of that variable with

[1] jack@jackhorne.net

the dependent. This leads to nonsensical/inactionable conclusions when treated in isolation. Should a company try to *reduce* satisfaction with the instruction manual to *increase* overall satisfaction due to a negative regression coefficient associated with the instruction manual? That hardly seems like a plausible strategy, especially if the pairwise correlation between the two variables is positive.

The above-described phenomenon is called multiple collinearity, or *multicollinearity* for short. Variables that are highly correlated with one another, and exert similar explanatory power on some dependent variable, are said to be collinear with one another. The problem is frequently exacerbated in survey data when variables are measured on the same rating scale and is further fueled by certain response-style biases. "Haloes" and "horns" effects are examples of such biases where a respondent is more likely to up-rate any features of a product that they are satisfied with, and to down-rate any features of product that they are dissatisfied with.

Multicollinearity is an early and well-documented problem in regression analysis (Ofir & Khuri, 1986). It results in increased variance around the regression estimates, potential sign reversals (i.e., counterintuitive effects), and coefficient instability. In practice, multicollinearity undermines the credibility of the analysis, and can make key drivers analysis in general a hard sell to stakeholders.

Several different approaches can be used to mitigate multicollinearity. Principal Components Regression (PCR), for example, derives orthogonal factors from predictor variables, and regresses those onto a dependent variable. The main disadvantage of this approach is that it results in the loss of the original variables, and potentially clouds interpretability.

Ridge regression is another method that seeks to reduce the variance around regression estimates by adding a *ridge parameter* (*k*) to the regression equation. Finding a value of *k* though that minimizes the variance around coefficients is not a simple task and is subject to researcher bias (Dorugade & Kashid, 2010).

Neither of these methods directly addresses the multicollinearity problem. A third approach seeks to do just that. Shapley regression is a modeling approach that determines relative importance of predictor variables by running and comparing every possible model among a set of predictors (Lipovetsky & Conklin, 2001; Conklin, Powaga & Lipovetsky, 2004). Shapley regression is derived from cooperative game theory where each player's input is determined as a combination of all possible combinations of the other players' input (Shapley, 1953). Lyon (2018) recently discussed some of the applications of Shapley analysis to market research data at the Sawtooth Software Conference.

As a worked example, say we wish to understand the effects of four predictors (v_1, v_2, v_3, v_4) on a single dependent variable. Shapley regression calculates all possible models involving those variables, compares them as shown in Figure 1, and computes averages across them to find the unique contribution of v_1 to R-squared of the whole model. It then repeats that process for v_2, v_3, and v_4.

Figure 1

Determining the Shapley Value for v_1 in a model using v_1, v_2, v_3, and v_4 as predictors.

Average the following effects, where f(X) is the effect of X predictors on a dependent variable.

R^2: $f(v_1) – 0$

R^2: $f(v_1, v_2)$ – R^2: $f(v_2)$

R^2: $f(v_1, v_3)$ – R^2: $\mathrm{f}(v_3)$

R^2: $f(v_1, v_4)$ – R^2: $f(v_4)$v

R^2: $f(v_1, v_2, v_3)$ – R^2: $f(v_2, v_3)$

R^2: $f(v_1, v_2, v_4)$ – R^2: $f(v_2, v_4)$

R^2: $f(v_1, v_3, v_4)$ – R^2: $f(v_3, v_4)$

R^2: $f(v_1, v_2, v_3, v_4)$ – R^2: $f(v_2, v_3, v_4)$

There are several appealing features to this variance partitioning approach. By averaging some statistic across all possible models involving a given predictor, we can capture the effect of that predictor independently from the effects of other related predictors. Shapley regression can also be applied to any type of regression model (e.g., linear regression, logistic regression, ridge regression, etc.). Shapley values sum to R^2 in a linear model (where R^2 is the usual objective function). An analogous objective function for logistic regression might be:

$(-2LL_{NULL}) – (-2LL_{model})$,

where all Shapley values would sum to this construct. For *any* objective function and any model type, the Shapley values sum to the objective function for the model with all variables included.

One disadvantage of Shapley regression is that it is computationally intensive. From Figure 1, we need to run 15 separate regression models to find the unique contribution of each of the predictors. (The same 15 regression models can be used to find unique contributions of any of the 4 variables, the results just need to be compared in different ways.) Running 15 models given 4 predictors is not computationally difficult, but the size of the problem increases exponentially as we add predictors. The number of models, given *p* predictors, is $2^p – 1$. Models with 10, 20, and 30 predictors require that we compute roughly 1,000, 1 million, and 1 billion separate regression models, respectively. For larger problems, we quickly run out of computing resources required for Shapley regression and it is not practical approach.[2]

[2] Sampling on the models and/or eliminating the most complex models can reduce computational needs while having very little effect on the Shapley values themselves.

Some prior research has been done in evaluating Shapley regression in the context of customer satisfaction (CSAT) work. Tang & Weiner (2005) pointed out that CSAT studies are often tracking studies and are repeated over time, say once a quarter. This creates the following challenge. Say we have four predictors, v_1, v_2, v_3, and v_4, and that all of these are highly correlated with each other and with overall satisfaction. In wave 1 the effect of v_1 turns out to have a relatively large regression coefficient while the effect of v_2 is minimal and has a regression coefficient near zero. We would conclude from this that v_1 is important and v_2 is not, even though both are substantially correlated with the dependent variable. This faulty conclusion is driven by multicollinearity among the variables in the model. Wave 2 of the same research might easily reverse the relationship between v_1 and v_2 because of very small changes in correlations, which would cause us to reverse our conclusion about which of the two variables are important.

Tang & Weiner found, using data from a commercial CSAT study that the average gaps among predictor relative importance across waves was smaller when using Shapley regression than it was when using simple ordinary least squares (OLS) regression. They further found that these gaps grew much faster as sample sizes decreased when using OLS compared to the Shapley regression-derived estimates.

What about prediction? Recall that we began this discussion with two goals in mind: *relative importance* and *prediction*. The output from Shapley regression is a statistic called the Shapley value. Shapley values are generated for each predictor in a model, and for linear regression are the part of R-squared that is uniquely due to each predictor. These are the values that we use to find relative importance. But prediction implies that we derive coefficients from a model that when applied to the data lead to estimates of the dependent variable. Lipovetsky (2006) developed a method to derive Shapley-adjusted regression coefficients by optimizing Shapley values for prediction and suggested using those coefficients in place of the OLS-derived coefficients.

Optimization problems are easily solved in R (R Core Team, 2021; Cortez, 2014; Mishra & Ram, 2019) and in other software. Example R code is given in the appendix to derive Shapley regression coefficients from Shapley values, and an R package is available on CRAN that calculates Shapley values and coefficients, given a dependent variable and a set of predictors (Horne, forthcoming, late 2021).

This paper tests the idea of using Shapley regression in a CSAT application. Does using Shapley-adjusted regression coefficients derived from the type of survey data gathered in CSAT studies meaningfully change predictions we make from the model compared to ones from OLS-derived regression estimates? If so, what are the implications of using either Shapley or OLS-derived coefficients in making predictions?

A Comparison Between Shapley and OLS Regression

A disguised commercial data set is used in this paper to illustrate Shapley value-derived relative importance and prediction from Shapley-adjusted regression coefficients and to compare both to the same metrics derived from linear OLS regression.

The data set (N=2000) pertains to a business-to-consumer delivery service, and consists of one variable measuring overall satisfaction, which was treated as a dependent variable, and six variables measuring satisfaction with specific aspects of the service, which were treated as predictors.[3]

Correlations between the predictors and the dependent were similar ($r = 0.651$ to 0.720). A linear OLS regression model (R^2: 0.595) on these same data resulted in regression coefficients that all matched the signs of the correlations, and most of which were statistically significant ($p < 0.05$). Relative importance derived from these coefficients, however, exaggerated what might have been expected from an earlier examination of the correlation coefficients. The ratio of the relative importance of the last two attributes, "delivery cost" and "on-time delivery," was 1.1 (18.7/16.8) when importances were derived from squared correlation coefficients and was 28.4 (41.1/1.4) when derived from OLS regression (see Table 1). Having to explain such a large discrepancy is a good example of where an analysis might lose credibility in relying on OLS coefficients to compare predictors. Delivery cost is likely not 30x more important than on-time delivery in determining satisfaction.

Table 1

Relative importance of predictors across three methods.

Predictor	Correlations		OLS		Shapley	
	r	rel.imp.	coef.	rel. imp.	value	rel.imp.
Accuracy of order (p1)	0.681	16.7	0.242	27.9	0.107	18.0
Quality of packaging (p2)	0.655	15.5	0.170	14.4	0.092	15.5
Friendliness (p3)	0.687	17.0	0.131	10.9	0.100	16.7
Tracking accuracy (p4)	0.651	15.3	0.094	4.3	0.084	14.1
Delivery cost (p5)	0.720	18.7	0.267	41.1	0.120	20.1
On-time delivery (p6)	0.683	16.8	0.053	1.4	0.093	15.6

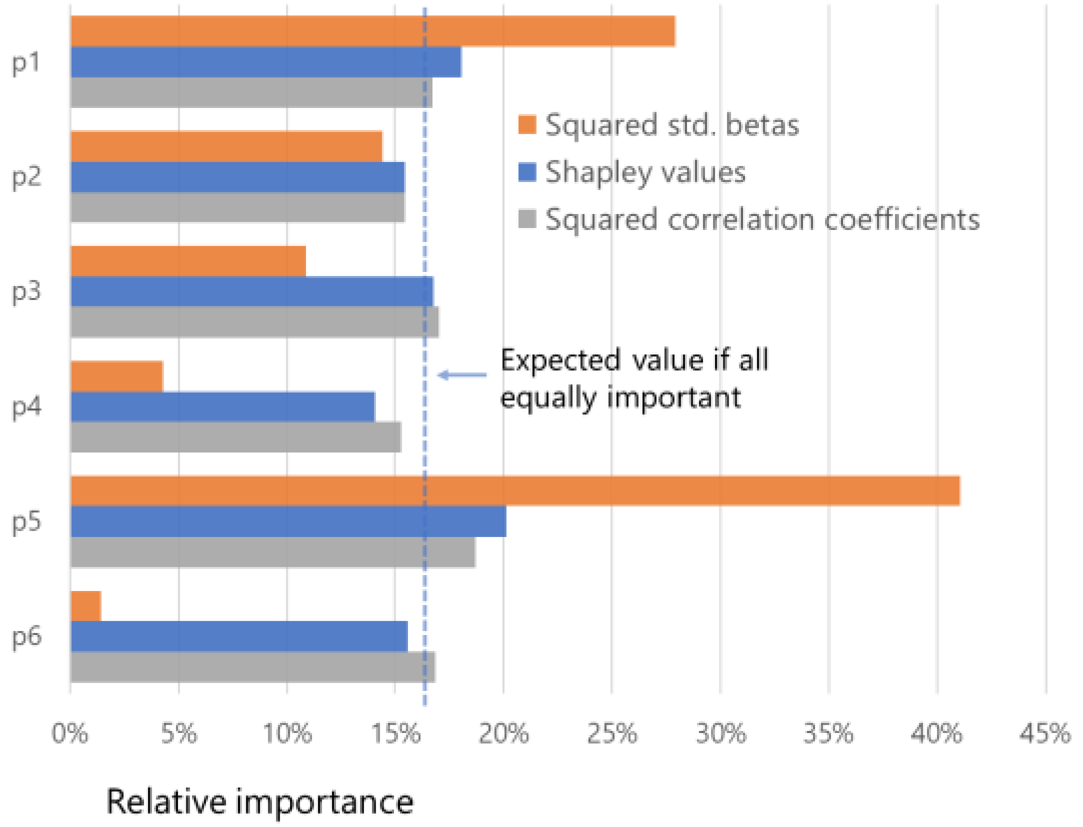
Shapley regression smoothed the relative importances compared to OLS regression. The ratio of relative importance from the last two attributes derived from Shapley analysis was 1.3 (20.1/15.6), which was much nearer what was found from the correlation coefficients than from the OLS regression coefficients. Figure 2 illustrates the differences in relative importance found from each of the three methods. It is clear from this demonstration that the Shapley values flatten relative importance compared to

[3] Descriptions of this data set are for illustrative purposes only. The actual attributes measured and the purpose for which these data were collected and analyzed has been disguised for the purposes of this paper.

OLS and reduce the distorting effects of relying on OLS regression coefficients to partition the unique effects of the predictor variables on the dependent.

Figure 2

Comparison of relative importance from three methods (1) squared standardized betas from OLS regression coefficients, (2) Shapley values, (3) Squared correlation coefficients. Relative importances are indexed to sum to 100 for each method.



A similar flattening occurs when comparing OLS regression coefficients to Shapley regression coefficients (i.e., those derived from optimizing the Shapley values for prediction; see Table 2). The mean regression coefficients for the two methods were 0.160 and 0.144, respectively, while the intercepts were 0.217 and 0.528. The smaller coefficients and larger intercept suggest a flattening of the slope of the regression line when in-sample predictions are made from the Shapley coefficients compared to the in-sample predictions made from OLS coefficients, which is what is in fact seen in Figure 3. Note that the predictions are nevertheless similar to one another, and that they differ nearly equally from the diagonal.

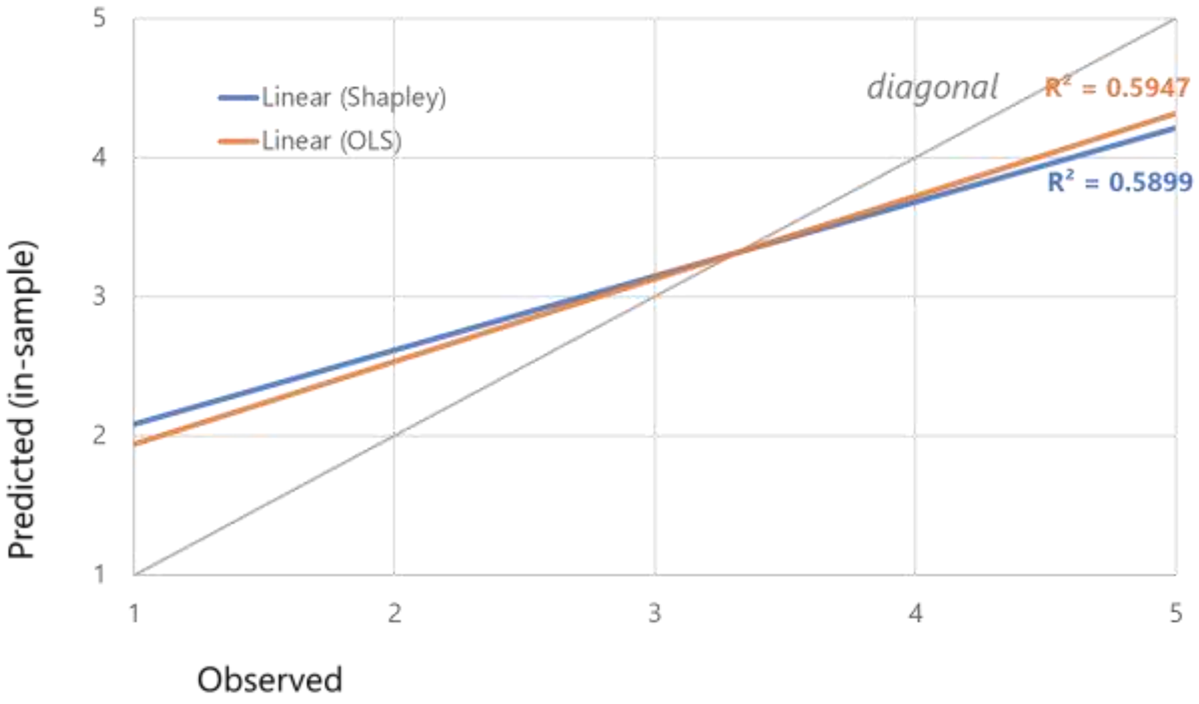
Table 2

Correlation and regression coefficients, Pearson r vs. OLS vs. Shapley.

Predictor	Pearson r	OLS Coefficient	Shapley Coefficient
Intercept	--	0.217	0.528
Accuracy of order (p1)	0.681	0.242	0.152
Quality of packaging (p2)	0.655	0.170	0.138
Friendliness (p3)	0.687	0.131	0.144
Tracking accuracy (p4)	0.651	0.094	0.131
Delivery cost (p5)	0.720	0.267	0.163
On-time delivery (p6)	0.683	0.053	0.138

Figure 3

Comparison of in-sample predictions made from OLS and Shapley regression coefficients.



Predicting holdouts using *k*-fold cross-validation tells a similar story. The *caret* package in R was used to generate the holdout samples (code available on request, k parameter=5 for all analyses using the method). The out-of-basket (OOB) fit was not quite as good for predictions made from Shapley coefficients as it was for the same predictions made from OLS coefficients, but the fit statistics were always within 1 standard deviation of each other for the two methods (see Table 3).

Table 3

Fit statistics.

Fit statistic	OLS	OLS	Shapley	Shapley
	In-sample	Out-of-sample*	In-sample	Out-of-sample*
MAE	0.684	0.687 (0.030)	0.701	0.702 (0.031)
RMSE	0.877	0.880 (0.024)	0.888	0.889 (0.022)
R^2	0.595	0.594 (0.021)	0.590	0.591 (0.021)

* mean (±sd) across hold-out samples.

Increasing the number of predictors or decreasing sample size[4] did not meaningfully change the fit statistics made from the two methods, though the variance around the fit statistics themselves did increase with decreasing sample size (see Table 4).

[4] The original data set consisted of 18 predictors, 6 of which were selected at random for the earlier analysis. Sample size was reduced in the same data set by randomly selecting cases.

Table 4

Mean absolute errors of predictions in actual dataset when varying number of predictors and sample size.

	OLS In-sample	OLS Out-of-sample*	Shapley In-sample	Shapley Out-of-sample*
# Predictors (all for n=2000 sample size)				
p=6	0.684	0.687 (0.030)	0.701	0.702 (0.031)
p=12	0.681	0.685 (0.033)	0.702	0.703 (0.035)
p=18	0.669	0.677 (0.037)	0.694	0.695 (0.035)
Sample size (all for p=6 predictors)				
n=200	0.679	0.691 (0.061)	0.692	0.694 (0.071)
n=500	0.695	0.709 (0.119)	0.724	0.727 (0.094)
n=1000	0.667	0.674 (0.025)	0.693	0.693 (0.027)
n=2000	0.684	0.687 (0.030)	0.701	0.702 (0.031)

* mean (±sd) across hold-out samples.

Simulating Different Correlation Structures

To explore beyond the actual data used as an example up to now, six simulated data sets were generated, each having different intercorrelation structures, using the *rnorm_multi* function in the *faux* package in R (code available upon request). Each data set consisted of nine variables, eight of which were treated as predictors and one as a dependent variable. Sample size was n=2000 for each data set, and correlation parameters are shown in Table 5. The correlation parameters were chosen to have both narrow and wide spreads of correlations, and to have low, medium, and high correlations among the variables.

Table 5

Mean absolute errors of predictions when varying correlation structures of simulated datasets.

	Mean absolute errors			
Correlation	OLS	OLS	Shapley	Shapley
parameters	In-sample	Out-of-sample*	In-sample	Out-of-sample*
Narrow spread				
1: r=0.15-0.25	0.961	0.965 (0.034)	0.963	0.966 (0.033)
2: r=0.45-0.55	0.713	0.717 (0.040)	0.717	0.718 (0.039)
3: r=0.75-0.85	0.372	0.374 (0.011)	0.419	0.419 (0.016)
Wide spread				
4: r=0.15-0.45	0.853	0.855 (0.014)	0.860	0.861 (0.013)
5: r=0.35-0.65	0.581	0.585 (0.025)	0.621	0.622 (0.022)
6: r=0.55-0.85	0.453	0.455 (0.013)	0.613	0.613 (0.026)

* mean (±sd) across hold-out samples.

OLS and Shapley regression coefficients were calculated for each data set and predictions were made from both methods using *k*-fold cross-validation (*k*=5). Fit statistics for the two methods and each data set are shown in Table 5. Fit was uniformly worse for Shapley predictions than it was for ones made from OLS coefficients. The difference between the methods was only outside 1 standard deviation when correlations were high or when there was a wide spread among the correlations.

Further examination of simulated data set 6 (wide spread, high correlations), which had the largest discrepancy in fit statistics between the two methods, shows that five of the eight predictors had different signs among the OLS regression coefficients than the associated correlation coefficients. The Shapley regression coefficients all matched the signs of the correlation coefficients. The absolute values of the Shapley regression coefficients were also much flatter (i.e., closer to zero) than those of the OLS regression coefficients (see Table 6).

Table 6

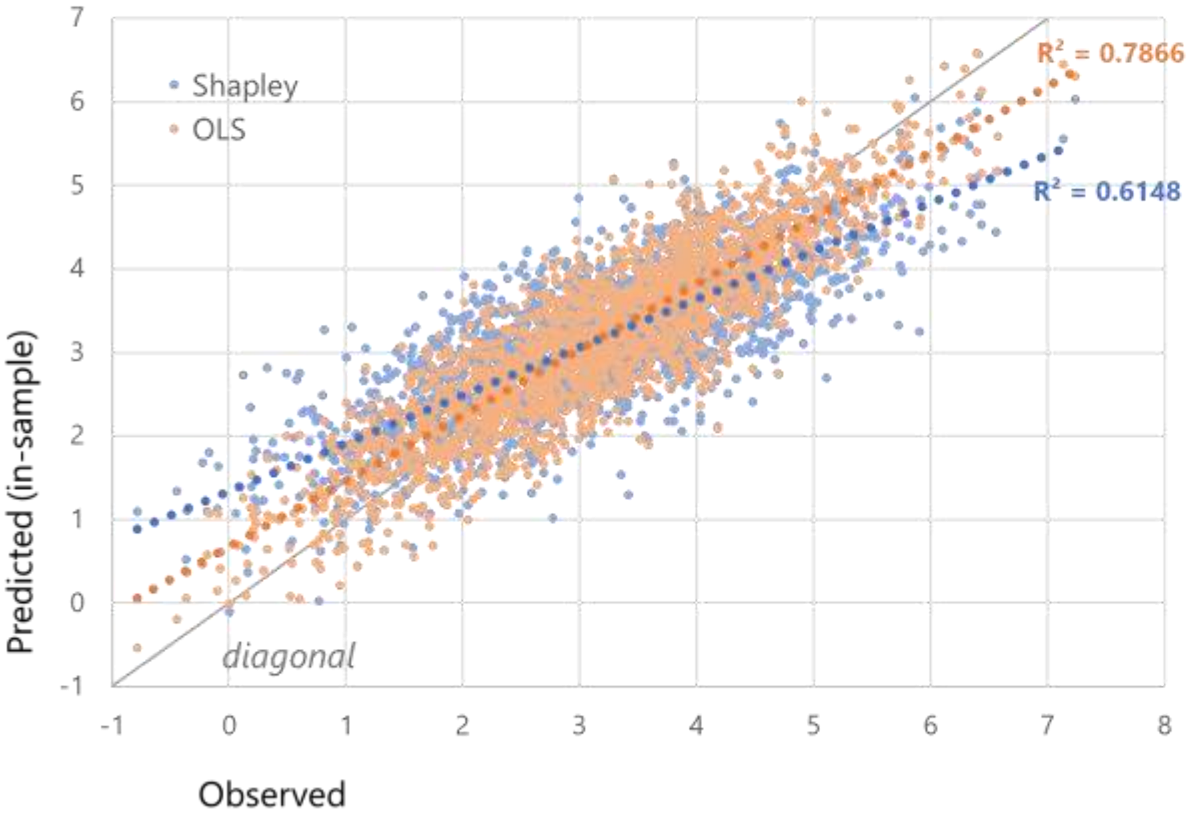
Correlation and regression coefficients, Pearson r vs. OLS vs. Shapley.

Predictor	Pearson r	OLS Coefficient	Shapley Coefficient
p1	0.571	-0.922	0.103
p2	0.585	0.595	0.079
p3	0.598	-0.033	0.070
p4	0.748	0.818	0.196
p5	0.551	-0.321	0.054
p6	0.522	-0.239	0.031
p7	0.710	1.005	0.190
p8	0.765	-0.162	0.162

This, as in the case of the actual data examined in this paper, again suggests a flattening of the slope of in-sample predictions made from Shapley coefficients compared to that of predictions made from OLS coefficients, which is in fact seen in Figure 4 for data set 6. In this case, the slopes for Shapley and OLS predictions were substantially different from one another, and predictions made from OLS coefficients better approximated the diagonal.

Figure 4

Comparison of in-sample predictions made from OLS and Shapley regression coefficients for simulated dataset 6.



Discussion

Survey-based customer satisfaction research is one of the most common applications of marketing research. Often, the customer satisfaction line item is the single largest budget item in marketing research departments, especially when customer satisfaction research is conducted as tracking studies over time. Market researchers and their stakeholders want and need to receive value for these large dollar projects.

When there are many components of customer satisfaction, each captured by a different survey question, management will be interested in identifying the “key drivers” of satisfaction from those components. In other words, they will want to know which of the many components of customer satisfaction they should focus on improving since there is likely not time or budget to improve all of them. OLS regression methods, be they linear or logistic, are frequently used to provide these insights. The present work has shown, as others have before, that use of these methods may distort relative importance of predictors, especially when there is a high degree of multicollinearity among those predictors. This in turn causes us to train our focus on some aspects, while diminishing other aspects that may in fact be equally important.

Shapley regression gets us better estimates of relative importance among predictors (i.e., more in line with what we'd expect from pairwise correlations), and as Tang & Weiner (2005) found, more stable estimates across waves of tracking studies, and with smaller sample sizes.

Alternatively, if we are seeking to use models of customer satisfaction to *predict* future values of customer satisfaction, it may be better to use an OLS model than it is to use Shapley-adjusted regression coefficients. This conclusion follows from the data-generating process involved in both approaches. At least from a linear modeling approach, the coefficients in an OLS model are derived from the data, using a closed-form process. The Shapley-adjusted regression coefficients on the other hand, are derived from an open-form optimization process. It makes sense, and this research tends to confirm, that there should be more noise around predictions made from an open-form process than ones made from a closed-form process.

In sum, the results presented here make the case for using Shapley values to determine relative importance among a set of predictors, but to not use Shapley-adjusted regression coefficients to make predictions from a model. The predictions from OLS models tended to do about the same or outperform the predictions made from Shapley regression in the data sets examined here. Of course, there are methods for making predictions among CSAT data other than OLS and Shapley regression. Some of these methods, such as random forests and other "learning" techniques, begin to get into the realm of machine learning. None of those methods were examined here but may be ripe for similar questions going forward.

Jack Horne

References

Cortez, P. (2014). Modern Optimization with R. Springer International.

Dorugade, A.V. & Kashid, D.N. (2010). Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences, 4(9), 447–456.

Conklin, M., Powaga, K. & Lipovetsky, S. (2004). Customer satisfaction analysis: Identification of key drivers. European Journal of Operational Research, 154(3), 819–827.

Horne, J. (2021). shapleyreg: Shapley regression for continuous and non-continuous variables. R package forthcoming on CRAN

Lipovetsky, S. (2006). Entropy criterion in logistic regression and Shapley value of predictors. Journal of Modern Applied Statistical Methods, 5(1), 95–106.

Lipovetsky, S. & Conklin, M. (2001). Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry, 17, 319–330.

Lyon, D. (2018). Shapley values: Easy, useful, and intuitive. In: Proceedings of the Sawtooth Software Conference, Orlando, FL. March 7–9, 2018. pp. 13–36.

Mishra & Ram (2019). Introduction to Unconstrained Optimization with R. Springer International.

Ofir, C. & Khuri, A. (1986). Multicollinearity in marketing models: Diagnostics and remedial measures. International Journal of Research in Marketing, 3, 181–205.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna Austria. URL https://www.R-project.org.

Shapley, L.S., A value for n-person games. In: Kuhn, H. and A. Tucker, Eds., Contributions to the Theory of Games, Princeton University Press.

Stratmann, W.C., Zastowny, T.R., Bayer, L.R., Adams, E.G., Black, G.S., & Fry, P.A. (1994). Patient satisfaction surveys and multicollinearity. Quality Management in Healthcare, 2, 2, 1–12.

Tang, J. & WeinerJ. (2005). Multicollinearity in CSAT studies. In: Proceedings of the 11th Sawtooth Software Conference, 83-90, San Diego, CA.

APPENDIX: EXAMPLE R CODE FOR OPTIMIZING SHAPLEY VALUES FOR PREDICTION

$$R^2 = 1 - (y - X\beta)'(y - X\beta)$$

$$= 2\beta'X'y - \beta'X'X\beta$$

$$= \beta'(2r - S\beta)$$

Given, $R^2 = \Sigma_j SV_j$

$$\beta_j(2r - S\beta)_j = SV_j, j = 1, \dots, n$$

then, function to be <u>minimized</u>:

$$F = \sum_{j=1}^{n} \left(\beta_j(2r - S\beta)_j - SV_j\right)^2$$

```
optFxn <- function(x, sv, corm) {

   svv <- as.vector(sv)
   cord <- corm[2:(length(svv)+1),1]
   corx <- corm[2:(length(svv)+1),2:(length(svv)+1)]
   ff <- sum((svv - (2 * x * cord) +
            (x * (rowSums(corx %*% diag(x))))) ^ 2)
   return(ff)
}
sval <- [vector of Shapley values]
corm <- cor(data)
scoeff <- nlminb(rep(1/length(sval), length(sval)), optFxn,
                 sv=sval, corm=corm)$par
```

AN INTEGRATIVE CONJOINT MODEL FOR COMPLEX PRODUCTS[1]

YICHUN MIRIAM LIU
OHIO STATE UNIVERSITY
JEFF D. BRAZELL
UNIVERSITY OF UTAH
GREG M. ALLENBY
Ohio State University

ABSTRACT

A complex product is made up of simpler, component products. A challenge in modeling demand for complex products lies in simultaneously studying features that affect component choice, and its resulting effect on marketplace demand. We propose an integrative model for studying complex products utilizing multiple conjoint exercises within a single structure. We illustrate our model using a conjoint study of demand for option packages when purchasing an automobile.

1. INTRODUCTION

Conjoint analysis is a stated preference research tool for measuring the value of product features. The attractiveness of conjoint analysis is that it imposes minimal assumptions about the structure of respondent utility, making it an ideal research tool for exploring the source of consumer preferences for product design and pricing decisions. Conjoint models are usually specified using dummy-variables for product attributes and their levels, allowing them to reflect an arbitrary utility function.

The most popular form of conjoint analysis is choice-based conjoint, where respondents express their preferences for products using discrete choices to indicate their most-preferred offering. Preferences are usually provided for marketplace offerings described by features of interest to firms to improve their offering. The purchase of an automobile involves choice from among different manufacturers with different types of engines (electric, gasoline, hybrid), drivetrains and option packages (e.g., driver assist, safety, luxury). Similarly, tourism packages may offer side excursions and upgrades, and mobile communication companies may bundle Wifi, television and cell phone services into a single offering.

The complexity of many product offerings makes it difficult to simultaneously study all of the features of a product (e.g., major drivers of a purchase decision and component-specific features), and studies are often commissioned to separately study various components of a product for component configuration. Automobile

[1] Generous financial support was provided by the Ford Motor Company, GRT00052940. The authors thank Anthony Bradfield, Peter Kurz, and Bryan Orme for providing helpful comments.

manufacturers, for example, use conjoint analysis to evaluate consumer preference for elements of safety, luxury or sport packages, and restaurants use conjoint to evaluate preferences for different menu items, seating options or aspects of ambience. However, these components are not available for purchase in the marketplace without the main, market-facing product. Results from the component study therefore need to be integrated with results involving the marketplace offering so that the effect of changes to component features on sales can be measured.

Ultimately, cars are sold in the marketplace, not safety packages, and the effect of changes to a product component need to be related to expected sales of the marketplace offering. In this paper, we show that integrating results from a focused, component conjoint study with the main, market-facing study needs to account for differences in the price levels of the component and the main products. Despite the flexibility of the assumed dummy-variable utility function used in a conjoint analysis, the utility for price is shown to exhibit diminishing marginal returns that leads to a systematic under-valuation of the component features when market and component prices differ. We find that economic measures of consumer preference for product features require adjustment beyond that suggested by linear models of demand and component integration.

We study consumer preferences for various option packages available in the automotive market using a unique dataset where respondents are asked to indicate their preferences across multiple conjoint exercises. Our results are consistent with evidence from the behavioral decision theory literature (Thaler, 1985) demonstrating that consumer price sensitivity is dependent on the base price of the item. This varying sensitivity becomes increasingly problematic for product design and analysis when the cost of component parts is increasingly different than the cost of the marketplace offering.

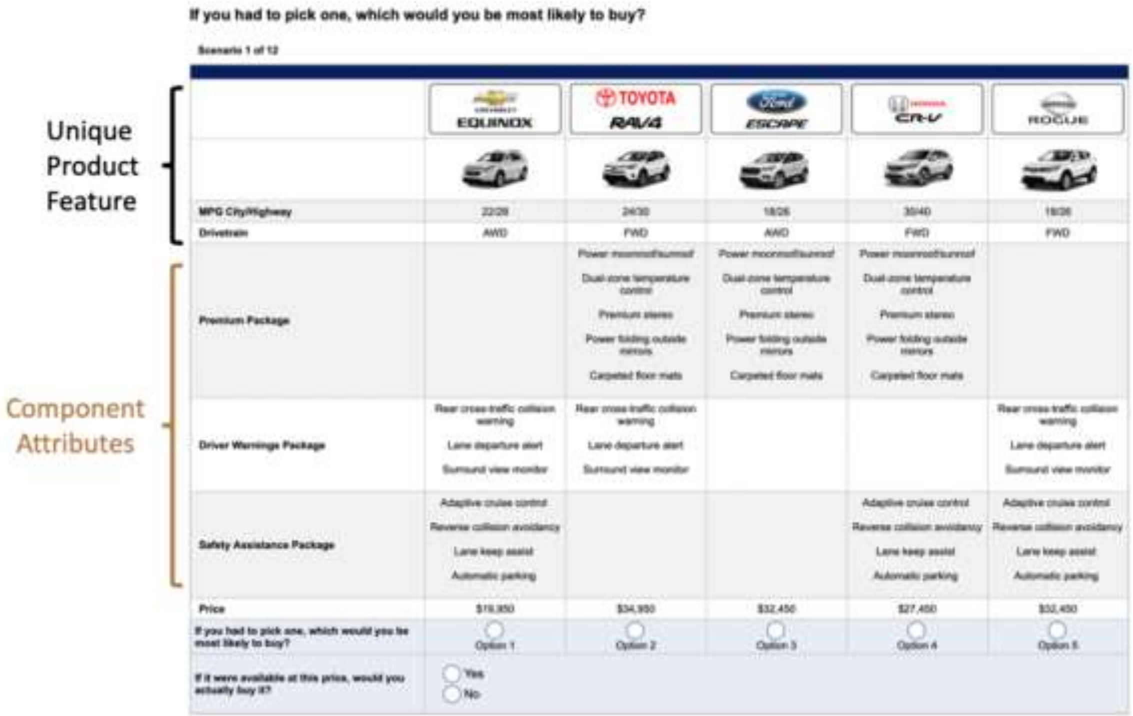
The organization of our paper is as follows. In the next section we illustrate our study design for evaluating complex product, followed by the discussion of our integrated model and alternative models. In Section 5 we report on estimation results, and in Section 6 we discuss implications for our findings. Managerial summaries are offered in Section 7.

2. Study Design

We investigate the effect of changes of complex product attributes on marketplace using a unique data set, in which respondents engaged in multiple conjoint exercises about the main and component features of Sport Utility Vehicles (SUVs). Respondents were qualified for inclusion in our study if they materially participated in a recent purchase of an automobile, or intended to make a purchase, where an SUV was considered. Data was obtained using a national Internet data provider (dynata.com) using a web-based survey.[2]

In the main, market-facing conjoint exercise, alternatives are described by four attributes that are unique to the marketplace offering (e.g., brand, gas mileage, and drivetrains), three component option package attributes (e.g., Premium, Driver Warnings, and Safety Assistance) with varying numbers of features, and a price attribute. The product features in the component option packages are either present or not present, and dummy variable coding was used to represent their presence. The price levels for the SUVs in the main conjoint exercise are approximately $25,000.

Figure 1: Main Conjoint Exercise



If you had to pick one, which would you be most likely to buy?

Scenario 1 of 12

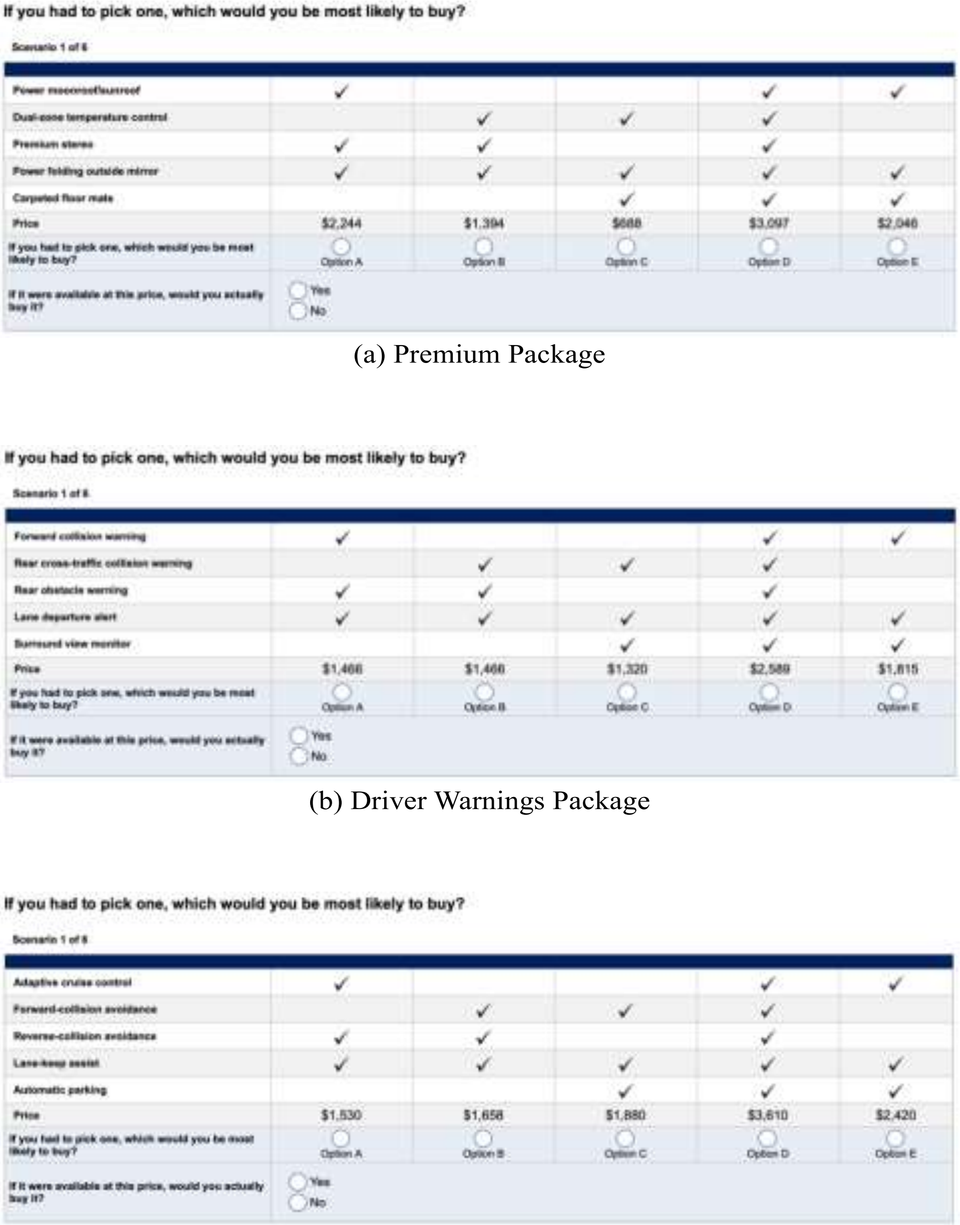
	EQUINOX	TOYOTA RAV4	ESCAPE	CR-V	ROGUE
MPG City/Highway	22/28	24/30	18/26	30/40	18/26
Drivetrain	AWD	FWD	AWD	FWD	FWD
Premium Package		Power moonroof/sunroof Dual-zone temperature control Premium stereo Power folding outside mirrors Carpeted floor mats	Power moonroof/sunroof Dual-zone temperature control Premium stereo Power folding outside mirrors Carpeted floor mats	Power moonroof/sunroof Dual-zone temperature control Premium stereo Power folding outside mirrors Carpeted floor mats	
Driver Warnings Package	Rear cross-traffic collision warning Lane departure alert Surround view monitor	Rear cross-traffic collision warning Lane departure alert Surround view monitor			Rear cross-traffic collision warning Lane departure alert Surround view monitor
Safety Assistance Package	Adaptive cruise control Reverse collision avoidancy Lane keep assist Automatic parking			Adaptive cruise control Reverse collision avoidancy Lane keep assist Automatic parking	Adaptive cruise control Reverse collision avoidancy Lane keep assist Automatic parking
Price	$19,950	$34,950	$32,450	$27,450	$32,450
If you had to pick one, which would you be most likely to buy?	○ Option 1	○ Option 2	○ Option 3	○ Option 4	○ Option 5
If it were available at this price, would you actually buy it?	○ Yes ○ No				

Figure 1 displays an example choice task for the main conjoint exercise, in which respondents are asked to select from among alternative SUVs for purchase. The component packages are comprised of up to five features that change across choice tasks. That is, the Premium Package is described in Figure 1 as having all available features, and this would change across choice tasks to allow estimation of all individual items within a package, or component.

[2] http://qa.interceptsurveys.net/Survey2320/default.aspx?RID=Test&VID=1&CID=1

In addition to the main conjoint exercise, our study design uses a corresponding number of component conjoint exercises to study the composition and value of components in complex products. Figure 2 shows the Premium, Driver Warnings, and Safety Assistance packages conjoint exercises. Alternatives in each component exercise are described by a set of package features that is common to the main conjoint exercise and a price attribute depending on the number of features listed, which varies across choice tasks. Prices in this task are smaller than that in the main conjoint exercise.

Figure 2: Component Conjoint Exercises



If you had to pick one, which would you be most likely to buy?

Scenario 1 of 8

Power moonroof/sunroof	✓			✓	✓
Dual-zone temperature control		✓	✓	✓	
Premium stereo	✓	✓		✓	
Power folding outside mirror	✓	✓	✓	✓	✓
Carpeted floor mats			✓	✓	✓
Price	$2,244	$1,394	$688	$3,097	$2,046
If you had to pick one, which would you be most likely to buy?	○ Option A	○ Option B	○ Option C	○ Option D	○ Option E
If it were available at this price, would you actually buy it?	○ Yes ○ No				

(a) Premium Package

If you had to pick one, which would you be most likely to buy?

Scenario 1 of 8

Forward collision warning	✓			✓	✓
Rear cross-traffic collision warning		✓	✓	✓	
Rear obstacle warning	✓	✓		✓	
Lane departure alert	✓	✓	✓	✓	✓
Surround view monitor			✓	✓	✓
Price	$1,466	$1,466	$1,320	$2,589	$1,815
If you had to pick one, which would you be most likely to buy?	○ Option A	○ Option B	○ Option C	○ Option D	○ Option E
If it were available at this price, would you actually buy it?	○ Yes ○ No				

(b) Driver Warnings Package

If you had to pick one, which would you be most likely to buy?

Scenario 1 of 8

Adaptive cruise control	✓			✓	✓
Forward-collision avoidance		✓	✓	✓	
Reverse-collision avoidance	✓	✓		✓	
Lane-keep assist	✓	✓	✓	✓	✓
Automatic parking			✓	✓	✓
Price	$1,530	$1,658	$1,880	$3,610	$2,420
If you had to pick one, which would you be most likely to buy?	○ Option A	○ Option B	○ Option C	○ Option D	○ Option E
If it were available at this price, would you actually buy it?	○ Yes ○ No				

(c) Safety Assistance Package

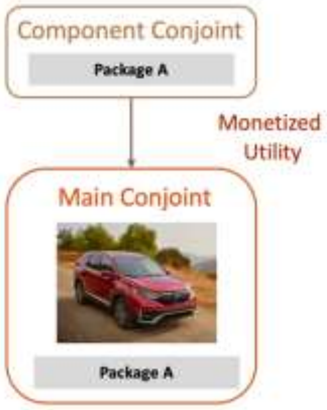
As complex products often have attributes numbers that exceed the suggested number of conjoint attributes and levels (Orme, 2020a), the specification of products in this way (e.g., multiple conjoint exercises) allows us to represent a large portion of the SUV market. In our study, respondents were asked to respond to one main conjoint exercise (twelve choice tasks), and three component conjoint exercises (six choice tasks each). Several screening criteria were introduced for data cleaning, a total of 547 respondents were available for analysis.

3. Our Integrative Model

Our approach to integrating the component and the main, market-facing conjoint exercise is through the monetized utility of the product components. We initially propose that the dollar value of changes in an attribute-level are the same across studies for an individual and allow the error scales to differ due to differences in exercise complexity. Figure 3 presents the concept of our integration mechanism in the context of one component option package, where:

1. Data from the component conjoint exercise are used to identify the monetized part-worths of component features within the package using the standard choice model with Sonnier et al. (2007) specification.
2. Given the monetized part-worths obtained from the component conjoint exercises, we calculate the monetized utility for the product component (e.g., Package A) presented in the main conjoint choice task.
3. As the choice alternatives in the main study is the combination of unique product features (e.g., brand, drivetrain, and gas mileage) and a product component composition (e.g., Package A), we account for the component by adding its monetized utility for the product component to the utility specification of the main study alternatives.

Figure 3: Concept of Proposed Integration Approach



The price coefficient in a choice-based conjoint analysis is equivalent to the reciprocal of the scale of the error term, which reflects the complexity of the choice task. Controlling for differences in task complexity is important because the main and component conjoint exercises can differ greatly in the number of included attributes (Swait and Louviere, 1993; Fiebig et al., 2010; Hauser et al., 2019). Our proposed model allows us to control for task complexity and account for heterogeneity. We also generalize our model to multiple components.

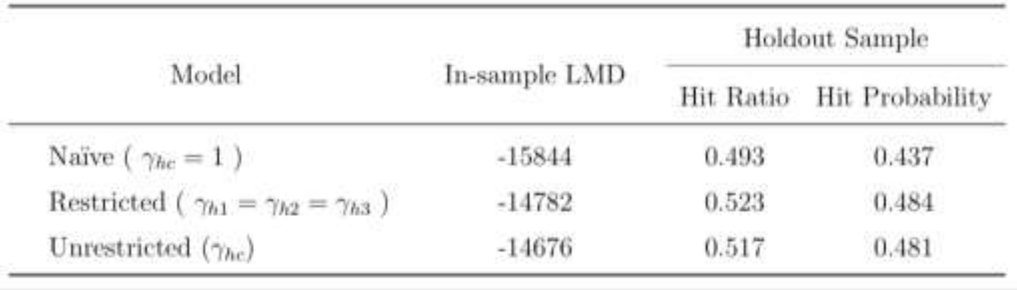
4. Alternative Models—Linear vs. Nonlinearity Price Effect

We examine three alternative specifications of our model to investigate how the main and component conjoint information integrate. A Naïve model integration assumes that the monetized utility of component features translates directly into the main conjoint study by adding the monetized package utility to the utility specification of the main study alternative. This integration approach implies a linear utility of money (e.g., remaining unspent budget). The use of a linear specification for integrating main and component exercises is expected to work well when the range of prices falls within a specific range. However, since the cost of a component can be much less than the cost of the market-facing offering, the integration of conjoint results across main and component exercises challenges the assumption of a linear specification and may lead to incorrect inferences about their market value. Our Restricted and Unrestricted model investigate the non-linear assumption of the remaining budget utility by allowing an adjustment to the monetized utility of the component product. This adjustment is captured by the parameter γ_{hc} introduced in our proposed model, where the subscript h and c denotes individual and component for analysis, respectively. The Restricted model assumes a constant adjustment across packages (e.g., $\gamma_{h1} = \gamma_{h2} = \gamma_{h3}$) whereas the Unrestricted model allows package specific adjustments (e.g., γ_{hc}). In short, the parameter γ_{hc} implies a non-linear specification of remaining budget utility if it does not equal one. We show that the linearity assumption (e.g., γ_{hc} = 1) is not supported when the general price levels are different, complicating inferences about the optimal composition and value of product components.

5. Results

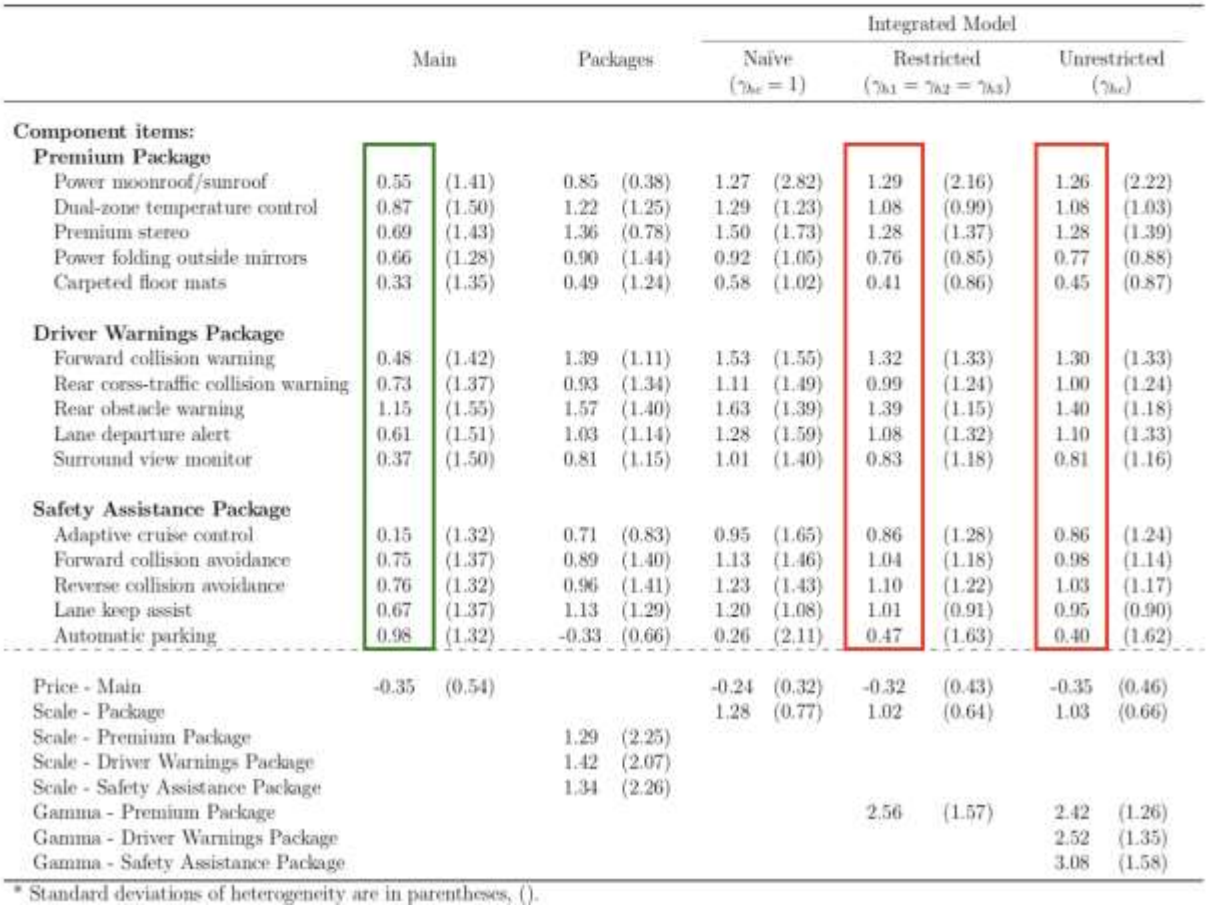
Table 1 shows the in-sample and predictive fit of the alternative models. The predictive fit measures indicate that the Unrestricted model fits the data best, although the Restricted model has about the same out-of-sample fit. Both models perform better than a Naïve model that impose no adjustment on component monetization for integration. These results indicate that utility for money (e.g., remaining unspent budget) is non-linear and that the integration of main and component conjoint exercises require adjustment beyond the simple monetization of utility by controlling for task complexity.

Table 1: In-sample and Predictive Fit



Model	In-sample LMD	Holdout Sample	
		Hit Ratio	Hit Probability
Naïve ($\gamma_{hc} = 1$)	-15844	0.493	0.437
Restricted ($\gamma_{h1} = \gamma_{h2} = \gamma_{h3}$)	-14782	0.523	0.484
Unrestricted (γ_{hc})	-14676	0.517	0.481

Table 2: Parameter Estimates



					Integrated Model					
	Main		Packages		Naïve $(\gamma_{hc} = 1)$		Restricted $(\gamma_{h1} = \gamma_{h2} = \gamma_{h3})$		Unrestricted (γ_{hc})	
Component items:										
Premium Package										
Power moonroof/sunroof	0.55	(1.41)	0.85	(0.38)	1.27	(2.82)	1.29	(2.16)	1.26	(2.22)
Dual-zone temperature control	0.87	(1.50)	1.22	(1.25)	1.29	(1.23)	1.08	(0.99)	1.08	(1.03)
Premium stereo	0.69	(1.43)	1.36	(0.78)	1.50	(1.73)	1.28	(1.37)	1.28	(1.39)
Power folding outside mirrors	0.66	(1.28)	0.90	(1.44)	0.92	(1.05)	0.76	(0.85)	0.77	(0.88)
Carpeted floor mats	0.33	(1.35)	0.49	(1.24)	0.58	(1.02)	0.41	(0.86)	0.45	(0.87)
Driver Warnings Package										
Forward collision warning	0.48	(1.42)	1.39	(1.11)	1.53	(1.55)	1.32	(1.33)	1.30	(1.33)
Rear corss-traffic collision warning	0.73	(1.37)	0.93	(1.34)	1.11	(1.49)	0.99	(1.24)	1.00	(1.24)
Rear obstacle warning	1.15	(1.55)	1.57	(1.40)	1.63	(1.39)	1.39	(1.15)	1.40	(1.18)
Lane departure alert	0.61	(1.51)	1.03	(1.14)	1.28	(1.59)	1.08	(1.32)	1.10	(1.33)
Surround view monitor	0.37	(1.50)	0.81	(1.15)	1.01	(1.40)	0.83	(1.18)	0.81	(1.16)
Safety Assistance Package										
Adaptive cruise control	0.15	(1.32)	0.71	(0.83)	0.95	(1.65)	0.86	(1.28)	0.86	(1.24)
Forward collision avoidance	0.75	(1.37)	0.89	(1.40)	1.13	(1.46)	1.04	(1.18)	0.98	(1.14)
Reverse collision avoidance	0.76	(1.32)	0.96	(1.41)	1.23	(1.43)	1.10	(1.22)	1.03	(1.17)
Lane keep assist	0.67	(1.37)	1.13	(1.29)	1.20	(1.08)	1.01	(0.91)	0.95	(0.90)
Automatic parking	0.98	(1.32)	-0.33	(0.66)	0.26	(2.11)	0.47	(1.63)	0.40	(1.62)
Price - Main	-0.35	(0.54)			-0.24	(0.32)	-0.32	(0.43)	-0.35	(0.46)
Scale - Package					1.28	(0.77)	1.02	(0.64)	1.03	(0.66)
Scale - Premium Package			1.29	(2.25)						
Scale - Driver Warnings Package			1.42	(2.07)						
Scale - Safety Assistance Package			1.34	(2.26)						
Gamma - Premium Package							2.56	(1.57)	2.42	(1.26)
Gamma - Driver Warnings Package									2.52	(1.35)
Gamma - Safety Assistance Package									3.08	(1.58)

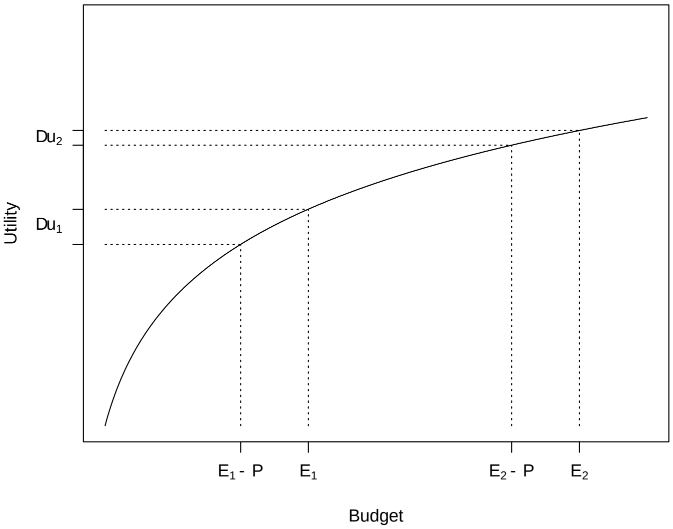
* Standard deviations of heterogeneity are in parentheses, ().

Part-worth estimates for the component product features are reported in Table 2. The first column reports the part-worth estimates from the main conjoint data alone, the second column reports parameter estimates from the three component conjoint models, and estimates from the integrated models appear on the right side of the table. We found that the error scale in the component package conjoint exercises is all estimated to be similar across the Main, Package, and Integrated Models. Despite of this similarity, we find the parameter estimates for the component package features in the main conjoint exercise to be smaller than those reported for the component package conjoint models and the integrated models. Implications of these findings are discussed in more detail below.

6. Discussion

Our analysis indicates that a market-facing study results in lower levels of price sensitivity than analysis conducted on the component features only. The reason is that sensitivity to changes in price is determined by the marginal utility of remaining unspent budget, which decreases as the level of expenditure increases (e.g., base price of the conjoint design).

Figure 4: Diminishing Price Sensitivity



Our analysis results indicate that even after controlling for differences in the error scale in the main and component exercises for choice task complexity, there remains a need to further adjust the influence of price on choices. We argue that this is evidence of a non-linear, concave utility function for the money (or price). Changes in price will have less of an effect on demand when the price level in conjoint exercise is large (e.g., cost of the market-facing offering) than when it is small (e.g., cost of the component product), and our analysis indicates that the effect can be substantial. In our analysis of the SUV market, there is a ten-fold difference in the prices in the main conjoint exercise than in the component conjoint exercises. We find that the adjustment needed is similar for all three component packages studied (i.e., Gammas in Table 2).

Our finding is consistent with behavioral decision theory (Thaler, 1985), suggesting that consumers react to percentage changes of price instead of levels of price. A classic example of people reacting to price changes is that more consumers would drive across town to save \$10 on a \$50 item than on a \$500 item, which is consistent with a diminishing marginal utility for alternative uses of money. Figure 4 illustrates this diminishing effect using a logarithmic function.

We examine two measures of economic value associated with the component packages—willingness to buy (WTB) and willingness to pay (WTP). Willingness to buy measures the predicted increase in share for a change in a product, holding prices fixed. Willingness to pay measures the expected increase in consumer welfare, measured monetarily, for a feature enhancement. WTP differs from estimates of monetized utility by accounting for the effects of competitive offers as discussed in Allenby et al. (2014).

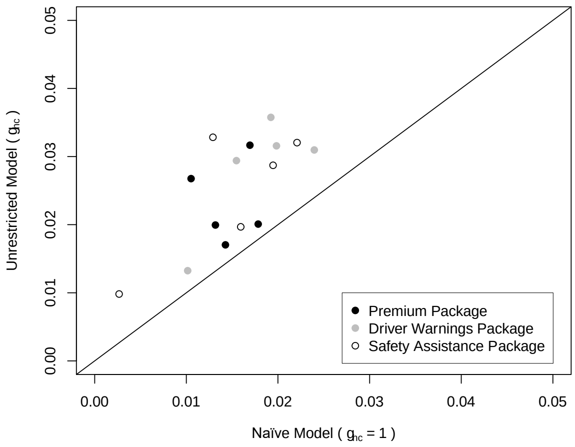
6.1 WTB

Figure 5 compares the effect of the component items on predicted changes in share (WTB) for the Ford Escape SUV. There are 15 points in the graph, one for each of the component items (e.g., Adaptive cruise control). We find that the estimates are all above the 45-degree line, indicating that the Naïve model under-predicts the influence of the component package items on share. The difference is, on average, equal to a 77.3% devaluation.

6.2 WTP

We investigate consumer willingness to pay (WTP) for package component items by comparing estimates from our model to the Manufacturer Suggested Retail Price (MSRP) of packages offered in the marketplace. Option packages are obtained from configuration webpages of automobile manufacturers from which we matched component items in our study. We examine three packages that align closely with our conjoint design: i) the Ford—Safe and Smart Package, ii) the Toyota—Premium Package, and iii) the Nissan—Driver Assist Package.

Figure 5: WTB: Predicted Change in Share for Ford Escape



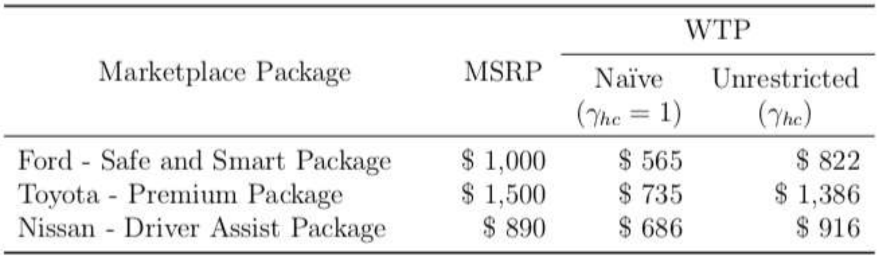
We measure WTP as the change in the monetized attainable utility for a change in a product offering.[3,4] We assume that the component items are available to all manufacturers, and therefore measure WTP as the decline in attainable utility to the consumer by its absence from the offering. Table 3 compares the MSRP to the WTP for the Naïve and Unrestricted models. We find that WTP estimates of the Unrestricted

[3] https://sawtoothsoftware.com/help/lighthouse-studio/manual/understanding-willingness-to-p.html

[4] https://sawtoothsoftware.com/resources/software-downloads/lighthouse-studio/version-history

model are closer to the MSRP than that of the Naïve model indicating the presence of a non-linear price effect.

Table 3: WTP for Marketplace Packages



Marketplace Package	MSRP	WTP	
		Naïve ($\gamma_{hc} = 1$)	Unrestricted (γ_{hc})
Ford - Safe and Smart Package	\$ 1,000	\$ 565	\$ 822
Toyota - Premium Package	\$ 1,500	\$ 735	\$ 1,386
Nissan - Driver Assist Package	\$ 890	\$ 686	\$ 916

7. Key Findings

This paper demonstrates the importance of conducting a market-facing analysis when evaluating component features. A market-facing analysis requires the inclusion of brand names and prices of products that are available for sale, in addition to the other product features that might comprise a component. Our integrated model provides a way to accurately study the composition and value of components in complex products, resulting in more accurate price/value information. We show that the integration of results from a focused, component conjoint analysis with a main, market-facing analysis needs to account for differences in the price levels of the component and main products. We find evidence of a non-linear pricing effect where component features are under-valued when the price of the component accounts for a small fraction of the total cost of a product. The key implications of our study are summarized as follows:

- A study of component features must include price of market-facing offer.
- Evidence of non-linear outside good and validate across three packages.
- Respondents are much less price sensitive when considering purchases in the main, as opposed to the component conjoint study.
- 77.3% devaluation of WTB without accounting the effects of non-linear outside good.
- WTP prediction of Unrestricted model is close to the MSRP.

This article is based on the presentation at the Sawtooth Software 2021 Conference. Additional details on our study and results can be found at: Liu, YiChun Miriam, Jeff D. Brazell, Greg M. Allenby (2021) An Integrative Model for Complex Conjoint Analysis, Available at SSRN 3696526, 2021
https://ssrn.com/abstract=3696526 or http://dx.doi.org/10.2139/ssrn.3696526

YiChun Miriam Liu

Jeff D. Brazell

Greg M. Allenby

References

Allenby, Greg M, Jeff Brazell, John R Howell, Peter E Rossi. 2014. Valuation of patented product features. The Journal of Law and Economics 57(3) 629–663.

Fiebig, Denzil G., Michael P. Keane, Jourdan Louviere, Nada Wasi. 2010. The generalized multinomial logit model: Accounting for scale and coefficient heterogeneity. Marketing Science 29(3) 393–421.

Hauser, John R, Felix Eggers, Matthew Selove. 2019. The strategic implications of scale in choice-based conjoint analysis. Marketing Science 38(6) 1059–1081.

Orme, Bryan K. 2020. Getting started with conjoint analysis: strategies for product design and pricing research. 4th ed. Research Publishers LLC, Manhattan Beach, CA.

Park, Young-Hoon, Min Ding, Vithala R Rao. 2008. Eliciting preference for complex products: A web-based upgrading method. Journal of Marketing Research 45(5) 562–574.

Liu, YiChun Miriam, Jeff D. Brazell, Greg M. Allenby. 2021. An Integrative Model for Complex Conjoint Analysis (April 08, 2021), Available at SSRN 3696526, 2021.

Swait, Joffre, Jordan Louviere. 1993. The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of marketing research 30(3) 305–314.

Thaler, Richard. 1985. Mental accounting and consumer choice. Marketing science 4(3) 199–214.

DISCUSSANT COMMENT ON "AN INTEGRATIVE MODEL FOR COMPLEX CONJOINT ANALYSIS"

JOEL HUBER
DUKE UNIVERSITY

Sawtooth Software users owe a debt of gratitude to Greg Allenby and his coworkers at Ohio State. More than 20 years ago Greg provided Sawtooth Software with the guidance to develop its Bayesian estimation of choice-based conjoint. More recently, following Greg's lead, Sawtooth Software is providing a willingness-to-pay estimate that adjusts simulated price effects for the outside good and reasonable competitive responses. In the presentation today, we can be particularly pleased by the exciting exploration which Greg Allenby, YiChun Liu and Jeff Brazell call an integrative model of complex conjoint analysis. It is complex in that almost 30 parameters are needed to account for automobile choice with market facing prices, joined with tradeoffs for features within groups of components. It is integrative because it uses careful econometric logic to merge these tasks. It is exciting because one set of conjoint exercises from 500 respondents can generate such insights on critical managerial decision-making.

The central finding from their study is that price sensitivity is greater for relatively inexpensive car components than for the complete market-facing package. This finding is an example of the way choices change with tasks that focus on or expand attention. Most choice conjoint exercises are constrained by restricting them to a particular context or to a small number of attributes, levels or alternatives. Those constraints tend to make the choices more consistent, have greater scale, and better able to predict choices within their domain. However, they also draw attention to attributes likely to be ignored in the marketplace.

The logical integration across studies that they produce is itself a remarkably clever feat. However the surveys would be valuable without the integration. Consider ways the main and component conjoint can be useful. The data from these studies can be linked to respondent characteristics to help determine the automotive needs of different segments. Simulations can then identify marketing and price changes that have promise to take sales from competitors. The component conjoint in itself is useful in helping to determine which features should be grouped into a defined clusters and which should be restricted to the market-facing decision. For example, it could determine effectiveness of the current practice in the Ford website to allow aftermarket carpeted floor mats for $200. Further, the component conjoint could be used to cluster features into components so that that people who value one feature are more likely to value the others.

The important point here is this study provides a coherent way to model a complex choice task. The empirical results may not be surprising, as it is expected that price sensitivity is greater for larger over smaller prices. We also know that price sensitivity is less for total prices compared with monthly payments, and that automotive test drives bring attention to personal and emotional qualities that can dominate choice. What we do not know is how to merge these different processes. Thus, Greg, YiChun and Jeff have shown us the impact of differential focus and how to deal with it. In that way their specific study becomes a role model for how to deal with complex choices, and we thank them heartedly for that.

Joel Huber