PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

May 2020

FOREWORD

These proceedings are a written report of the twenty-first Sawtooth Software Conference, held in San Diego, California, September 25-27, 2019. Two-hundred attendees participated.

The focus of the Sawtooth Software Conference continues to be quantitative methods in marketing research. The authors were charged with delivering presentations of value to both the most sophisticated and least sophisticated attendees. Topics included conjoint analysis, surveying on mobile platforms, MaxDiff, market segmentation and classification, experimental design, and the perils of establishing causality in observational data.

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, and David Lyon.

Sawtooth Software

May, 2020

CONTENTS

SUMMARY OF FINDINGS

The twenty-first Sawtooth Software Conference was held in San Diego, California, September 25-27, 2019. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2019 Sawtooth Software Conference Proceedings.

A Comparison of PC and Mobile Interviewing Modalities (Deb Ploskonka, Karlan Witt, Cambia Information Group): Deb and Karlan noted that the increase in mobile survey completions has been accompanied by higher rates of breakoff rates for mobile survey takers. To help reduce breakoff rates among mobile survey takers, the authors tested two question types among US and Japanese respondents where they see concerning breakoff rates: unaided brand awareness and grid-style brand rating questions. For unaided brand awareness, they tested a version of the survey that showed 15 blank entry boxes versus five blank entry boxes that expanded with extra entry boxes as respondents filled in the blanks. There was very low abandonment among the US-based respondents, with no difference between the versions. Japanese respondents had higher abandonment for the 15 blank entry box approach. For the grid portion of their experiment, Deb and Karlan tested three approaches: standard grid, scroll approach with items broken into separate ratings questions, and a “freeze” version which was the same as scrolling but left the header portion of the question frozen at the top of the screen. The results were inconclusive regarding which method resulted in the lowest dropouts and most consistency between laptop and mobile data. For future research, Deb and Karlan noted that there are other styles of mobile grids (accordion and progressive grids) that could be tested.

***A Researcher's Guide to Studying Large Attribute Sets in Choice-Based Conjoint** (Megan Peitz, Numerious, Mike Serpetti, Dan Yardley, Gongos): Choice-based methods have become dominant in our industry, yet no clear answer has emerged regarding which of the many choice-based approaches a researcher should use as the number of attributes increases beyond about six. Megan and co-authors designed an experiment involving choice of smartphones by real respondents to compare Partial Profile CBC, Adaptive Choice-Based conjoint (both partial-profile and full-profile variants), and Full-Profile CBC for attribute lists involving 10, 15, and 20 attributes. They found that for studies involving 10 attributes, the methods were on parity with each other in all respects (holdout predictability, length of survey, data quality, and respondent perception). As the number of attributes increased to 15+, the results suggested that partial-profile ACBC has a number of advantages over the other techniques. The authors also pointed out the large difference in the predicted None rate between ACBC and CBC.

*Best Presentation based on audience voting.

What Do South African Medical Students Value in a Rural Internship (Maria Jose, University of Cape Town, South Africa): To address inequality in access to healthcare among rural populations in South Africa, effective health worker recruitment initiatives to these areas are vital. Maria conducted a Choice-Based Conjoint study (CBC) among final year medical students in South Africa to identify their most valued rural health facility attributes for internship placement. Results, estimated by hierarchical Bayes estimation, demonstrated that the attributes of physical safety at the facility, availability of basic resources, degree of practical experience gained at the facility and the availability of personal protective equipment against tuberculosis exposure were more valued than the current recruitment initiatives of housing provision and rural allowance. Simulations demonstrated that even if rural allowance were to be increased by 20% it would not off-set the disutility of working at an unsafe facility, nor one which did not provide personal protective equipment to prevent tuberculosis exposure. The study thus provides evidence for policy makers to invest more in the working conditions of South African rural healthcare facilities' infrastructure, security and supply chains to the benefit of their patients and staff alike.

Leadership Qualities: Preferences from the Millennial Generation (Ronald Mellado Miller, UVU, Christina A. Hubner, Sawtooth Software Cray Daniel Rawlings, and Maureen Andrade, UVU): Ronald and his coauthors drew upon leadership characteristics gathered from business researchers and evolutionary psychology, then surveyed business students to see which characteristics they would prefer to have in a CEO. Using MaxDiff, respondents marked the CEO characteristics that were most and least important to them when considering a future job. The authors applied Latent Class Multinomial Logit analysis to discover two distinct groups, a "Sensitive Group" and an "Achievement Group," that preferred contrasting traits in a CEO. Ronald and his coauthors also found Gender differences in what respondents preferred in a CEO. The study suggests that previous research on preferences in leadership may need modification for the Millennial demographic. Studies like their can help businesses and other employers understand how to best attract potential candidates by emphasizing particular CEO characteristics as well as hiring CEOs that appeal to their future workforce.

Virtual Reality Meets Traditional Research: or the Reality behind Virtual Reality Enhanced Interviews (Alexandra Chirilov, GfK): Alexandra described how VR is becoming more widely available to consumers and at the same time more industries are using the power of VR. For market researchers, VR offers a new set of tools to provide a richer, more immersive experience that allows us to test products in a realistic way that better recreates the environment in which the customers experience products. Applications include store layout tests to car clinics to product development. Alexandra conducted an experiment comparing VR-enabled CBC with CAPI CBC. The VR-programmed survey used HTC Vive headset and controllers. VR respondents took a little over 6 minutes on average in a VR training room prior to taking the survey. Once in the survey, VR respondents needed only 40% more time to answer the questionnaire compared to the CAPI one. Mode of interviewing (CAPI vs VR) didn't affect the conjoint importances, though the utility scores for levels within attributes did change in a few instances. Preference for model and wheel types differed. Holdout predictability was very similar between VR and CAPI. Alexandra concluded that the use of VR engaged the respondents more fully, creating a more satisfying survey environment than CAPI, which in the end translated into a better quality interview.

Too Much Information?: The Curious Case of Augmented MaxDiff (Jackie Guthart, Curtis Frazier, and Raman Saini, Radius Global Market Research): When the number of items in a MaxDiff study is in the range of 30-60, Jackie and her co-authors prefer to use Augmented MaxDiff which involves asking preference questions outside the MaxDiff section and augmenting the MaxDiff data with that external information. These external questions can be in the form of Q-Sort or ranking questions of items chosen "best" in the MaxDiff exercise. How to add the augmented information to the MaxDiff data was the crux of the authors' investigation covering three augmented MaxDiff studies involving real respondents. If respondents ranked 10 items in the external information, this could involve coding all 45 inferred comparisons in the most aggressive treatment. That most aggressive augmentation leads to the highest withing-respondent fit in HB estimation, but may be overfitting. They also tested sparse coding of the augmented information. Based on different treatments across their three MaxDiff studies, the authors recommend against exhaustively specifying the inferred pairs for HB estimation. Rather, they recommended a moderate amount of augmentation, such as a differential approach that specifies relatively more augmented pairs involving the best-ranked items and fewer augmented pairs involving the worst-ranked items.

Can We Use RLH to Assess Respondent Quality? (Marco Hoogerbrugge and Menno de Jong, SKIM): Marco reviewed previous recommendations for using HB's RLH fit statistic to identify bad or random respondents. He noted that some bad respondents can actually get lucky and obtain reasonable looking RLH scores. Therefore, Marco stated that using HB RLH scores isn't the best approach to identify bad respondents. He introduced a new approach that involves adding purely random responders to a CBC dataset (at a rate of about 1/6 the number of real respondents in the original real dataset), then running a high-dimensionality Latent Class MNL solution on the combined data set (i.e. involving 20 or even 30 groups). Those groups that are majority composed of random responders are marked as "bad" latent class groups. All real respondents with higher likelihood of belonging to any "bad" groups compared to other groups are to be discarded as "bad" or near-random responders.

Bandit MaxDiff: The Effects of Design Parameters on Hit Rates in Diverse Datasets (Nico Peruzzi, elucidate): Bandit MaxDiff has emerged as a good method for identifying the top (best) items from very large lists (100+ items). Compared to a fixed sparse design, Bandit achieves similar hit rates on top items while using approximately one-fourth the sample size. In this paper, Nico explored the effects of dataset characteristics and Bandit parameter adjustments (supported by Sawtooth Software's implementation) on top item hit rates. The number of items sampled for each respondent to evaluate across the MaxDiff questions has little effect on hit rates, whether set at 20, 30 or 40 items. Reducing the number of items shown to each respondent could help reduce cognitive load. The number of sets shown to each respondent does have a negative effect on top item hit rates, however, this negative effect is less in the case of fewer items under study (60 vs 120), less error (noise) in the dataset, and when sample size is large. Reducing the number of sets asked of any one respondent reduces survey length and can be considered if other parameters are optimally set.

When the focus is on identifying top items, Sawtooth Software's default for Thompson sampling featuring 5/6 of the items selected for a respondent to evaluate (relative to 1/6 selected based on items seen fewest times by previous respondents) performed better than the less aggressive setting of half-and-half. Nico concluded with recommendations for approaching a Bandit MaxDiff study based on dataset characteristics and Bandit parameters to have the best shot at achieving high hit rates for top items.

Trees, Forests and Situational Choice Experiments (Keith Chrzan, Sawtooth Software, and Joseph Retzer, ACT-Solutions): Keith and Joe described situational choice experiments and compared using polytomous multinomial logit and several machine learning methods (decision trees, two types of random forests and a gradient boosting method called CatBoost) to analyze them. Across 10 empirical data sets they found that CatBoost had superior predictive validity, as it had the highest 10-fold cross validation rates. In terms of explanation, however, Keith and Joe suggested that polytomous MNL and decision trees best allow researchers to understand and communicate the respondents' decision process (both enable the analyst to supply the end user with an easy-to-use Excel simulator), with trees having the additional benefit of providing an easy-to-communicate visualization of the decision process.

Examining the No-Choice Option in Conjoint Analysis (Maggie Chwalek, Roger A. Bailey, and Greg M. Allenby, Ohio State University): Maggie and coauthors stressed that for valid economic interpretation a conjoint analysis must include, at a minimum, each alternative's brand name, prices, and an outside "no-choice" option. Respondents use the no-choice option to indicate that some other offering not included in the choice set is preferred to those included in the choice set, or that it is better for them to hold onto their money and not make a purchase. Thus, selecting the no-choice option assumes that respondents have some knowledge of the prices and features of the real marketplace. Maggie and coauthors conducted a choice experiment to examine the effect of providing respondents with information about the prices and features of tooth whitening products. They found that conjoint estimates are surprisingly robust to the information provided about existing marketplace options. They concluded that screening and qualifying respondents based on product participation are sufficient for identifying qualified candidates who are aware of marketplace prices such that providing additional pricing and attribute information will only minimally affect part-worth estimates.

Modelling Stockpilable Product Purchase Decisions Using Volumetric Choice Experiments (Richard T. Carson, University of California, San Diego, Towhidul Islam, University of Guelph, Jordan J. Louviere, University of South Australia): In categories such as canned tuna, buyers often decide how many units of a specific good to purchase rather than simply deciding whether to purchase it or not. Richard and coauthors stressed that there is much more information in such count data than in traditional discrete choices. They designed and implemented a volumetric choice experiment (VCE) defined by price, brand, size, and other attributes. The VCE experimental design allowed for identification of over 100 own- (brand by size) and cross-price elasticities via a multilevel mixed-effects negative binomial regression. Their VCE design allowed for the estimation of a very rich deeply parameterized model, with a focus on being able to provide own-price elasticities for each brand-size combination and a complete set of cross-price elasticities. They concluded that modeling differences in own-price elasticities and brand-specific constants provides a much richer picture about what is happening in the market than models that only allow for one price effect.

Conjoint Meets AI (Peter Kurz, Stefan Binner, bms marketing research + strategy): AI and Artificial Neural Networks (ANN) have been applied to utility estimation for discrete choice experiments, but HB has generally proven more robust. Peter and Stefan decided to investigate whether ANNs could be used in the design of discrete choice experiments, especially for alternative-specific designs. Theoretically efficient designs feature orthogonality and level balance, but they often lead to choice tasks featuring dominated concepts and lack of utility balance. The authors argue, however, that either no utility balance or too high of utility balance are both bad. Other complex designs such as line pricing (where multiple SKUs' prices move in tandem) are also out of the range of traditional orthogonal plans. The statistical goal of their design search procedures was to handle complex design needs while resulting in utility estimates for purely random data that were as close as possible to zero. To that end, they employed the Softmax procedure from ANN that does MNL estimation. To find optimal designs, ANNs need a large number of versions & answers to select the subset of versions and answers that minimize the loss function. They tested their ANN approach to experimental design compared to traditional designs as generated by SAS. Respondents found the ANN designs more realistic, with more products available that they would like to buy. Hit rates of holdout tasks for the ANN designs were also superior to the SAS designs. Peter and Stefan concluded that AI-based design techniques appear to deliver more stable and better results, although two empirical studies are not enough to conclude that they are always superior.

Predicting the (Unobserved) Predictable: The Use of Deep Learning in Wave Studies for Market Research (Tom Gardner, Michelle McNamara, Adelphi Research): In pharmaceutical primary market research, conducting multiple wave studies can be costly as it often requires recruiting physicians, whose time is expensive. Clients still need to gather this information, therefore Tom and Michelle explored the question of whether there was a way to optimize the survey time by predicting the more predictable attributes. They first identified which attributes to use using Principal Components analysis. Using the data from two waves they trained a Convolutional Neural Network to predict four attributes (as well as fitting a linear model for comparison). These models were then used to predict the responses of a third wave. The results showed that at the aggregate level, the Convolutional Neural Network was extremely precise and only deviated from the actual results by a small margin. At the respondent level, Convolutional Neural Network performed better than the linear model at predicting an external variable. The implication of this approach is that fewer questions could be asked of the respondent, with little loss in precision.

Can I Reduce the Number of Conjoint Tasks and Still Get Good Quality Data? (Chris Moore, Ioannis Tsalamanis, Ipsos MORI – UK): Chris and Ioannis began by noting how the trend is for online surveys to take less time, especially as more people complete surveys on mobile. Thus, they examined ways to deal with even shorter CBC surveys than typical practice. They tested whether common data imputation routines (model-based and distance-based) could help in the case of sparse data by imputing answers to CBC questions that are supplied by different respondents who completed the same version (block) of CBC questions. Chris and Ionnis re-examined different real CBC datasets, by purposefully deleting respondent answers to subsets of the CBC tasks to create different degrees of missing data. They found that current practice with HB estimation on sparse data performed better than using HB estimation on sparse CBC datasets that have first been enhanced by data imputation.

Combining Choice-Based Conjoint and Dynamic Choice Models for More Accurate Forecasting (Faina Shmulyian, SKIM USA): Some purchases involve well-known products such as dish detergent. Other purchases involve complex products where the buyer's understanding of the product and its benefits evolves over time (high tech and innovative products). Faina's presentation investigated how to measure the impact of innovation (advertising) and imitation (word-of-mouth in social networks) on the diffusion process (market penetration into a consumer base) for a complex and innovative product: DNA testing kits.

Data Fusion: A Flexible HB Template for Modeling Structures across Multiple Data Sets (Kevin Lattery, SKIM Group): In our age of expanding data we are more likely to find ourselves with two or more sources of data, Kevin explained. When we need to make sense of these multiple data sources in relation to each other, this is what we call data fusion. Kevin described three general approaches to data fusion: 1) Two-Stage Linkage, 2) Data Augmentation/Stacking, and 3) Complete Structural Model/Probabilistic Programming. The last model is the most complex and requires specialized programming. Kevin tested the structural model versus the simple approach for choice data for a few real data sets. For a simple anchored MaxDiff Kevin did not see much benefit in using the structural model over Sawtooth Software's approach of stacking. However, when fusing MaxDiff plus rating scales, the structural modeling showed significant gains. Kevin demonstrated another example involving fusing CBC and MaxDiff data, where there was modest improvement for the structural model.

Segmenting Choice and Non-Choice Data Simultaneously: Part Deux (Tom Eagle, Eagle Analytics of California, Inc., Jay Magidson, Statistical Innovations, Inc.): Tom and Jay began the discussion with a reminder to HB users that it isn't considered best practice to run cluster analysis on HB estimates (from either CBC or MaxDiff) even if the utilities are first normalized. Rather, latent class choice models provide a more sound approach. Recently, Latent Gold software released an update that allows scale-adjusted latent class (SALC) to be used with one or more continuous variables. That advance also leads to the opportunity to see if latent class clustering could be applied to continuous HB utility data. They concluded that the SALC model can produce meaningful segments not only when based on raw Best-Worst choices (the best practice/gold standard approach), but also when used to cluster on HB utilities derived from the Best-Worst choices. Dave Lyon's discussion of this paper clearly demonstrates the issues with segmenting raw HB utilities. Another thing Tom and Jay investigated was a new option available in Latent Gold for weighting the impact of variables differentially in latent class clustering.

Understanding Consumer Preferences: A Comparison of Survey and Purchase-Based Approaches (James Pitcher, Bradley Taylor, and Dan Kelly, GfK): James and co-authors reported on one of the largest and most comprehensive research studies to compare attribute importance, brand preference, and price elasticities between conjoint and POS (Point of Sale) data across 15 technology and durables product categories and 7 countries. They found that although attribute importance and brand preferences are similar, there are large differences in price elasticities between the POS and conjoint models due to the differing ways in which they measure consumer preferences and due to some weaknesses in the POS data (namely, multicollinearity). Conjoint measures a theoretical preference, one that is not influenced by external market factors, whereas the POS data takes into account the in-store realities and how these affect the purchase decision. POS models may therefore be beneficial when tactically modelling specific market scenarios as they are closer to market realities. However, POS models cannot be used for new product development, testing new features or new prices, or measuring feature preference. The best approach in such circumstances is still conjoint, they concluded.

Maximizing the Impact of OOH (Outdoor) Advertisement Using Discrete Choice Modeling and Text Analytics (Rajat Goel and Rachin Gupta, StatWorld Research Solutions): In this modern era of competition, advertising plays an important role in the success of a brand. Out-of-home (OOH) advertising is considered one of the most important modes of advertising. Rajat and Rachin showed an innovative research application that used Discrete Choice Modeling along with Text Analytics to create outdoor ads with maximum impact and likeability. While Discrete Choice Modeling was used to pick the various visual elements of the advertisement, the wording was driven by text analytics. The research allowed Rajat and Rachin to create an advertisement which not only addressed limitations of previous approaches, but also had the potential to have maximum likeability and impact among the consumers. They expressed the opinion that this kind of approach can be applied to any case where one can clearly identify and separate the elements that make up an advertisement.

Using Adaptive Choice-Based Conjoint Analysis to Unravel the Determinants of Voter Choices (David Bakken, Foreseeable Futures Group, Gretchen Helmke, University of Rochester, and Mitchell Sanders, Meliora Research): David and co-authors used Adaptive Choice-Based Conjoint analysis (ACBC) to understand the impact of a candidate's party, specific policy positions, and orientations toward democratic principles on individual choices between candidates. According to the presenters, conjoint analysis has only recently become popular in political science. They cited a few recent studies involving CBC, but posited that in the presence of screening rules or other heuristics (such as must-have attribute levels), the pairwise designs of these studies could fail to capture the marginal effects of some attributes. To extend the literature, they conducted an online survey with a general population sample of 1005 US adults leveraging ACBC. They also included a MaxDiff section that involved twelve named candidates to gauge voter preference for the candidates. Their two main questions with the study were, a) is ACBC an appropriate and perhaps better approach than other conjoint methods for understanding voter preferences and predicting their electoral choices? and b) to what extent will voters trade off democratic values in order to maintain partisan loyalty. On the second question, David and co-authors' findings were similar to those reported by previous researchers applying conjoint analysis to political choices. Respondents do appear willing to trade off their preferences for democratic values in order to choose a candidate that reflects their policy and party preferences. Regarding the first question, they found only small benefits for using ACBC instead of CBC for conjoint research in the political arena.

The Challenge of Identifying Causality in Observational Data (Ray Poynter, The Future Place/Nottingham Trent University): There has been an explosion in the amount of observational data available to decision makers and research. However, there are challenges in the use of observational data, Ray emphasized, such as making the link between correlation and causality, survivor bias, homophily, and combinatorial effects. Ray noted that observational data should be embraced and utilized, but the challenges should be recognized and dealt with. Despite the many potential advantages of observational data, Ray expressed that controlled experiments are still seen as the gold standard. Identifying the counterfactual is a key step in assessing causality. A control cell in an experiment is a counterfactual, as well as matching people who have and haven't seen a social media campaign creates a counterfactual. And, as a general best-practices recommendation, researchers should seek to minimize the number of independent (or predictor) variables, Ray suggested, and to maximize the number of independent observations.

A Comparison of PC and Mobile Interviewing Modalities

Deb Ploskonka
Karlan Witt
Cambia Information Group

Abstract

Over a third of Americans (37% according to Pew) now go online mostly using a smartphone, which explains the increase in mobile survey rates researchers have experienced. This increase suggests a need and movement toward mobile-enabled surveys. Mobile-enabled surveys, however, often lack the ability to meet survey takers' expectations. Many respondents note layout difficulties such as having to zoom in to the question text, scrolling to see the full range of answers and not being able to immediately see the question due to screen size.

Based on these limitations, survey providers notice higher breakoff rates among mobile respondents. In an attempt to reduce survey breakoff among mobile users, Cambia Information Group tested the two question types where we see higher breakoff rates among respondents in the U.S. and Japan: unaided brand awareness and grid-style brand rating questions. These were tested in various formats on both desktop and mobile platforms across these two cultures.

This paper includes salient background references for this topic, question construction and formatting details, analysis of the resulting data, our recommendations for best practices going forward, and proposed extensions of this research.

Introduction

In a recent Sawtooth Software Conference, there was audience discussion about best practices for the presentation of survey questions given the prevalence of mobile devices. Historically, surveys had been programmed for someone to take on a PC, and there were concerns expressed both about the quality of the experience someone would have taking that survey on a mobile device, as well as whether or not we should program a mobile version to look and feel more like a native mobile experience rather than a desktop experience being viewed on a phone.

In the ensuing discussion, there was little agreement as to approach and even less data to inform best practices going forward. Cambia Information Group has tested the options articulated by the audience and the results are shown in this paper. Through our investigations and audience discussion at the 2019 conference, we learned that this issue continues to be an area of concern for several reasons:

- Many organizations have legacy tracking surveys where implementing revised mobile approaches might impact trendability;
- Evidence shows a large percentage of surveys are still not mobile-friendly, potentially causing non-response bias or simplification techniques among mobile users; and

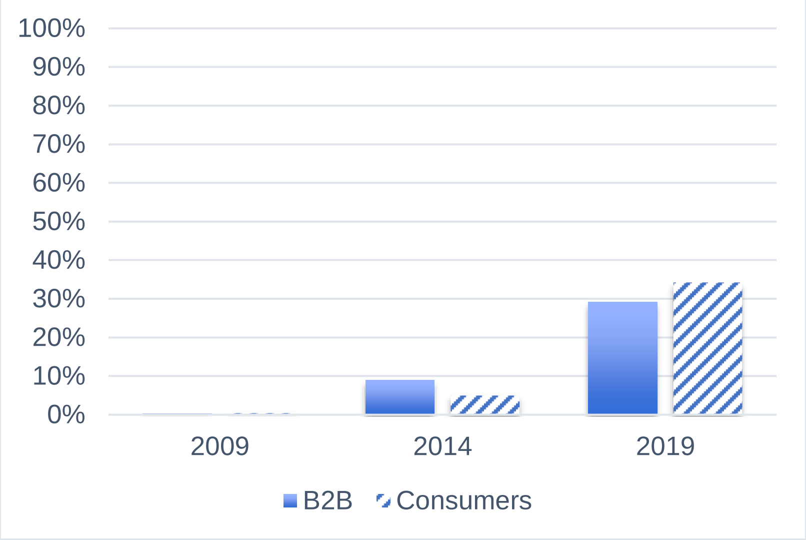
- There remains debate as to whether the look and feel should be consistent across mobile and desktop platforms or optimized for each.

While most of the Sawtooth Software Conference is focused on expanding on the most sophisticated areas of conjoint and modeling, there is a critical need to ensure quality data is collected to serve as a basis for this advanced work. This paper aims to serve in this capacity.

Background: The Need for Mobile

Depending on the types of respondents you survey, mobile may seem like more or less of an issue to you. We conducted this research-on-research based on trackers we run globally over time showing significant increases in mobile survey completion for both business and consumer audiences. Figure 1 below shows the increase in mobile completions for both B2B and B2C studies.

Figure 1: Percent mobile of an international multi-audience tracker over time (consumer audience: panel; B2B: lists)



We reached out to some panel partners who confirmed they are seeing high mobile survey rates as well.

- Symmetric reports 25% on mobile with 10% on tablet.
- Dynata cites 30% on mobile, though some studies are not available to or are discouraged for mobile.
- Lucid's figures top 50% as some countries' Internet access rates are driven by increasing access to smartphones.

Figures 2 and 3 below, from data provided by Dynata through September 2019, show we've reached a tipping point where mobile and desktop survey starts are roughly even, with a striking trend for mobile usage both domestically and internationally.

Figure 2: U.S. Survey Starts by Device

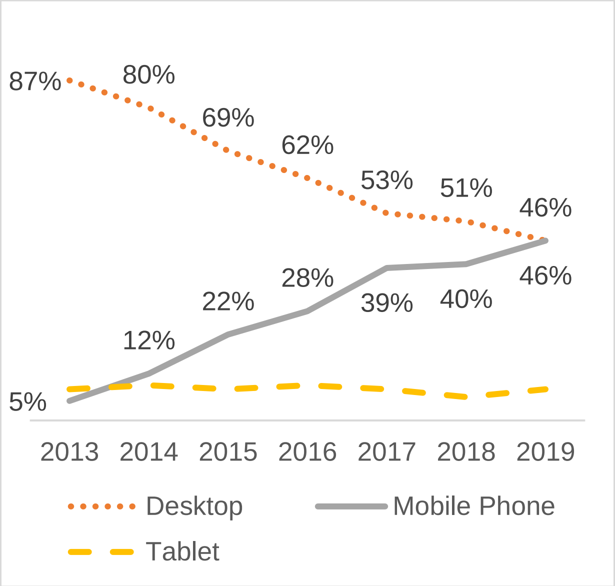
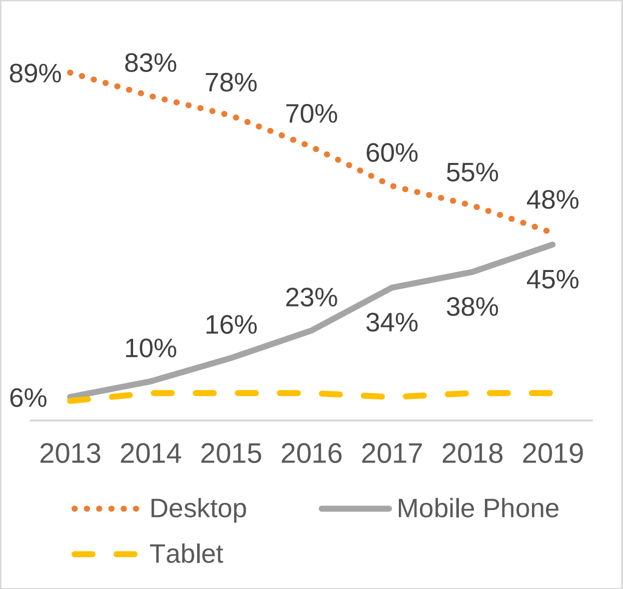


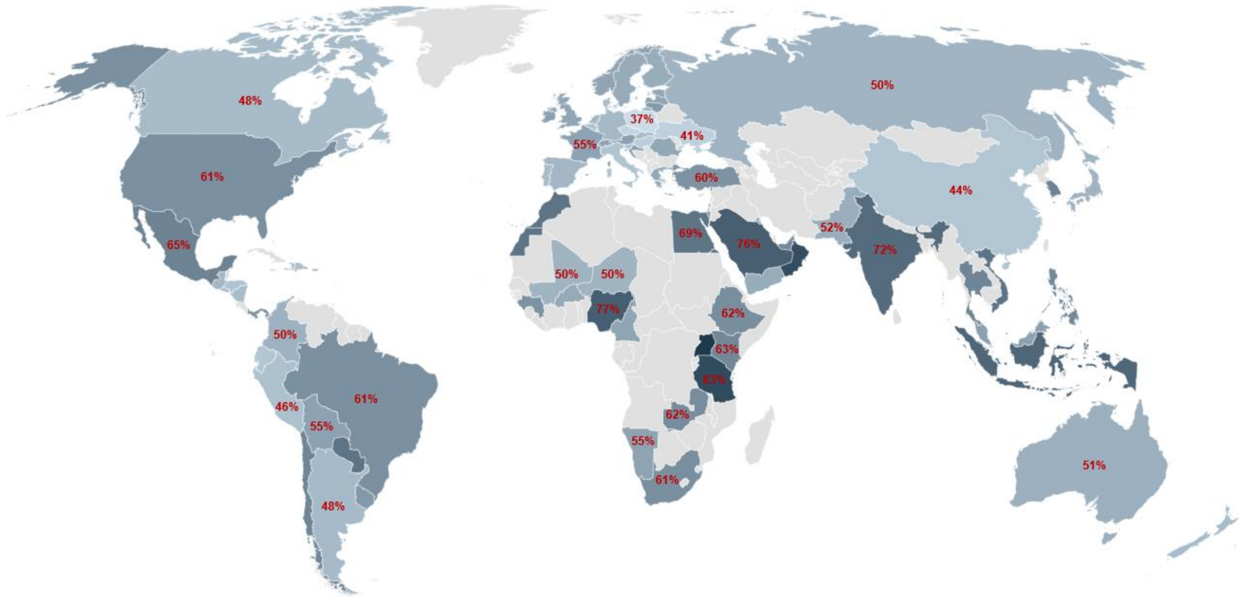
Figure 3: Global Survey Starts by Device



To reiterate the obvious, survey presentation on mobile devices will have a huge impact on the data that is collected.

Mobile usage continues to vary by country. Lucid's survey starts in 2019, seen in Figure 4, go from a low of 37% in Poland to a high of 89% in Uganda.

Figure 4: Survey Starts on Mobile by Country (Lucid)



While mobile is present in a big way worldwide, there has not been a uniform approach for how to deal with it.

One country where Cambia surveys often, India, shows 72% mobile starts. Both McKinsey (McKinsey Global Institute, 2019) and Kantar (Kantar IMRB, 2019) have reported huge growth of internet penetration in India, primarily via smartphone as data plans, and the phones themselves, have dropped in price. In the U.S., broadband acquisition in some segments has dropped as smartphone capabilities increase and allow users full internet accessibility. 37% of Americans now go online *mostly* using a smartphone, says Pew recently (Pew Research Center, 2019).

The State of Mobile Surveys Today

When we decided to write this paper, friends and colleagues began forwarding us specific notes they were getting when trying to take surveys on mobile devices. And there were MANY.

Some were mobile-friendly, and just recommended that for a better mobile experience use the landscape view to maximize real estate on the screen. Examples include:

- Please rotate your device to landscape mode for this question.
- We noticed that you've started this survey on a tablet or smartphone. To optimize your survey experience, we recommend using landscape view.
- It appears you're taking this survey on a smartphone. We suggest rotating your smartphone to landscape position (i.e., horizontal) for a better survey-taking experience.

Some said you are going to have a bad experience on a mobile device, we recommend you use a desktop PC instead:

- While this can be taken on a mobile device, we recommend a desktop for ease of selection.
- Please note that since this survey contains multiple grid-questions and it can be taken with greater ease on a desktop, tablet or laptop/computer than on a cell phone.
- Not all of our clients' surveys are designed for mobile devices. If you choose to use a mobile device, you may not have an optimal experience.

The most extreme was literally not allowing mobile surveys to participate, even though the survey content had nothing to do with being a PC user:

- There are elements of this survey that are incompatible with smartphones […] please close your browser and switch over to a desktop/laptop computer to complete this survey.

So, who are we missing when our surveys are not mobile-friendly?

Who you miss is skewed by a number of variables such as generation. And younger folks can be harder to get to respond to a survey under any circumstances, but telling them they can't do it on their phone can really skew your underlying sample and introduces unknown non-response bias.

We asked our friends and colleagues who had been given the "you will have a bad experience on mobile, please switch to a PC now" type message if they had indeed switched. While this is a convenience sample, zero of the folks who received these types of messages went back to take the survey later when in front of a desktop PC. We did some research and found a study that examined switching behavior in this situation.

Research published by Peterson (Peterson, Griffin, LaFrance, & Li, 2017) indicates these messages encouraging switching to a desktop have a non-meaningful impact on switching (0.9% switch compared to 0.4% of the control group). So, in practice, if we enforce a desktop-only survey, there will be systematic non-response bias in our data.

In our own data, we also noted differences beyond generation that could also have a material impact on your resulting data, including gender, ethnicity, life stage, and those active on social media. Figures 5 and 6 show behavior by generation in U.S. and Japan roughly the same, with survey-taking behavior on phone increasing the younger the generation of the respondent.

Figure 5: U.S. Percent by Generation 2017+

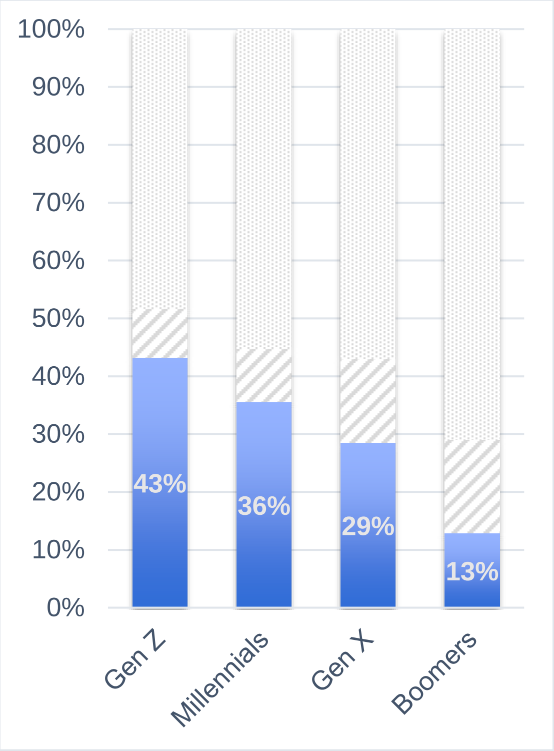
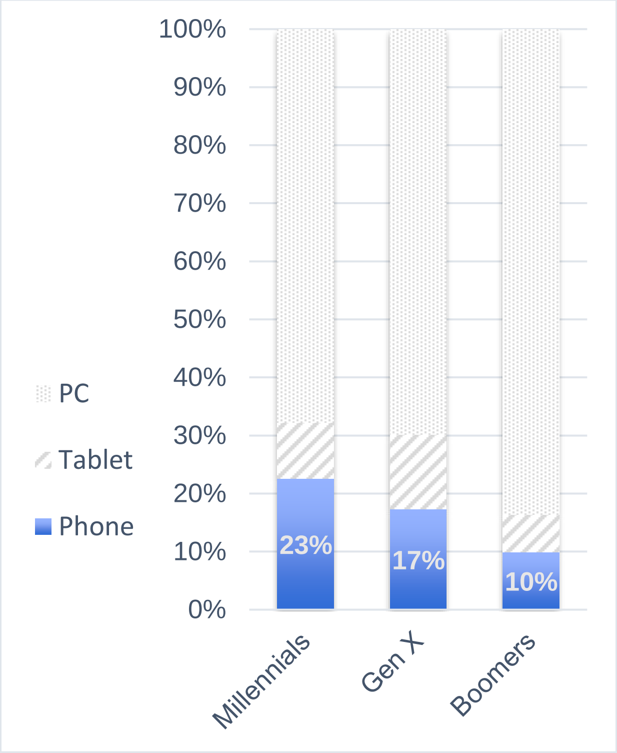


Figure 6: Japan Percent by Generation 2017+



So, the respondents you lose are likely to be systematically different from those who complete the survey.

Other Considerations

There are some straightforward best practices related to question layout for mobile devices:

- Respondents need to:
  - Read question without zooming or horizontal scrolling
  - Record answers without zooming
  - See the question immediately

The second topic is a more important one. This exact topic was the one raised at the last Sawtooth conference. It centers around having identical look and feel across desktop and mobile devices vs. optimizing for each. Is it best for the *data* to:

1. Have a single mobile-enabled layout for both PC and mobile,
2. Have a single mobile-*optimized* layout for both PC and mobile, or
3. Choose a layout that best matches the device? This last option would prioritize the *respondent* experience.

As Ray Poynter has said in his book on the topic, "mobile surveys can be thought of as existing on a continuum from intentional at one end, for example where the survey was designed for mobile and where the participants were selected because they had a mobile device, through to 'unintentional mobile' where the survey was designed for PC and unexpectedly participants complete it on their mobile." (Poynter, Williams, & York, 2014)

The argument for aligning the stimuli is to remove the impact that the type of device the survey is taken on from how they are answering the question. The downside is these are usually desktop-PC optimized and run into all the problems we've been talking about.

The argument made for optimizing for *each* platform to maximize the respondent experience cited the different look and feel and even interactivity we all experience on our phones vs. desktop PCs. You can hover your mouse over something on a PC, but unless you specifically program "touch" vs. "tap" for phones and tell the respondent about it, they won't have hover-over access.

Respondents using mobile devices may have certain expectations. We all use mobile versions of websites that automatically detect the device we are using to access the site and tile or lay out content differently to optimize the user experience. Forcing mobile-savvy respondents to take a survey designed for a desktop PC might increase abandonment rates, or survey "breakoff."

Our Challenge: Reducing Breakoff in Mobile

In the early days of mobile devices, most online surveys were still taken on traditional desktop and laptop PCs. While some testing was conducted to confirm functionality when viewed on a phone, little guidance was available to change the way in which questions were asked. We had a long-term global tracking study that was launched before mobile devices were this prevalent, so in fairness would consider this particular survey "mobile enabled" but not necessarily "mobile optimized."

As we watched mobile completion rates rise, we saw disproportionate breakoff rates for two question types:

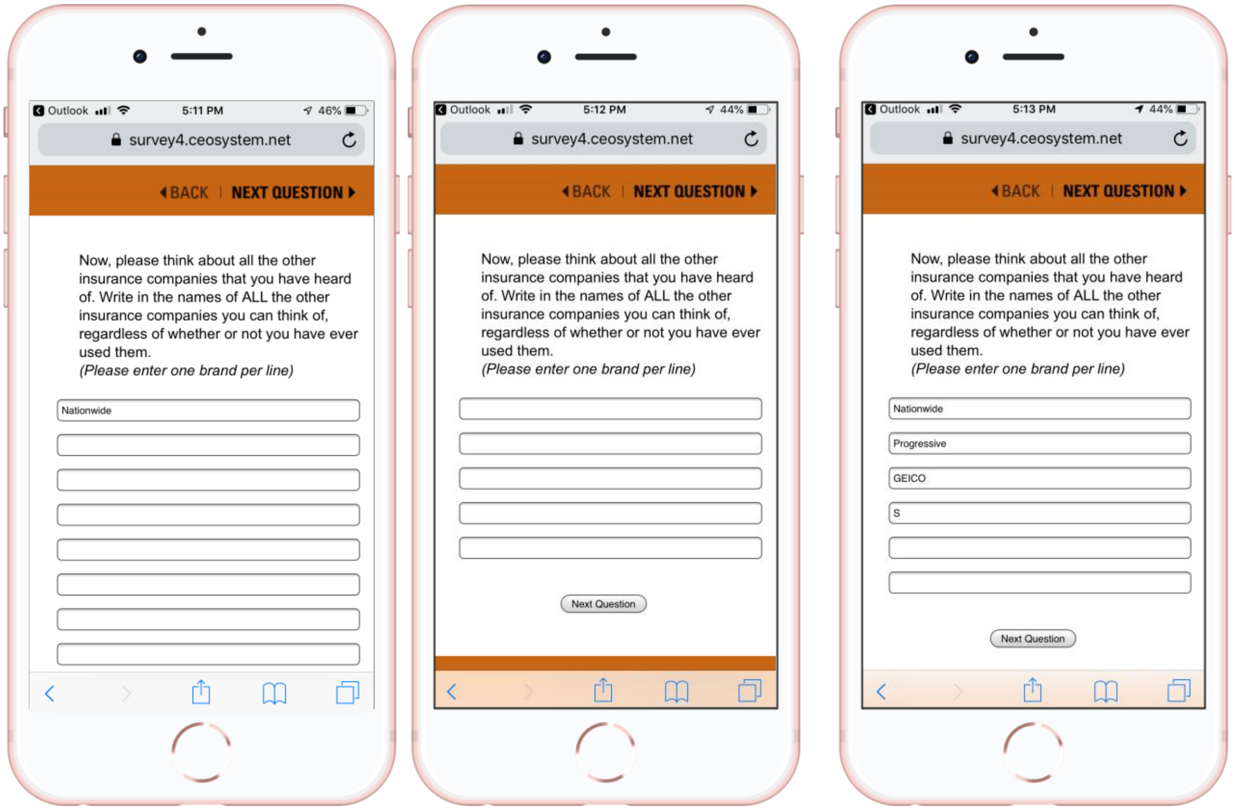
- Unaided awareness
- Grid-style brand rating questions

As well, the new version of Lighthouse Studio had a new mobile construct for grids we were anxious to test. So, we conducted research-on-research to examine the versions we had historically run against some new alternatives to test several of our hypotheses.

Unaided Awareness Test

The subject matter for our test comes from a long-running tracker we were conducting in the insurance industry, in as many as 25 countries worldwide. On the left, below, is how we had been asking an unaided brand awareness question, following a top of mind single open-ended question, with 15 lines shown. After grids, this question type had the second highest breakoff. We hypothesized it was worse for mobile due to the idea of typing so many lines on a small screen with a small keyboard.

Figure 7: Unaided Awareness Test Design



Now, please think about all the other insurance companies that you have heard of. Write in the names of ALL the other insurance companies you can think of, regardless of whether or not you have ever used them.
(Please enter one brand per line)

Nationwide

Now, please think about all the other insurance companies that you have heard of. Write in the names of ALL the other insurance companies you can think of, regardless of whether or not you have ever used them.
(Please enter one brand per line)

Now, please think about all the other insurance companies that you have heard of. Write in the names of ALL the other insurance companies you can think of, regardless of whether or not you have ever used them.
(Please enter one brand per line)

Nationwide
Progressive
GEICO
S

In the middle and at right above, in Figure 7, is our test version, displaying only five lines to start, and then progressively adding lines to always have two blank lines as respondents filled in their responses, with a maximum of 15. We intended this to be a more engaging and also less overwhelming experience, particularly on mobile.

With the study inspiring this test conducted internationally, we did not want to arrive at US-centric conclusions, thus half our data was collected in Japan. We have equal sample size per control and test, per PC and mobile. As you saw above, incidence of tablet starts is low, and as some research-on-research groups it with PC and some with mobile, we opted to set it aside. Results are controlled for demographic differences, given the obvious differences in demographics on who might respond on which device.

Sample sizes per cell per country per device ranged from 152 to 168. Data were collected from August 21 to August 28, 2017.

Hypotheses

This test had two hypotheses:

1. Seeing 15 open-ended prompts may be overwhelming. By only showing five initially, and then successively more, the experience would be more engaging and less overwhelming, **reducing break-off**.
2. As only one-quarter of respondents gave more than five answers in the preceding tracking study, the quantity and quality of responses **would not significantly change**.

As is well known by the industry, response rates have been dropping over time across all modalities. By increasing the engagement on a question that provokes a higher level of dropouts we wanted to improve response rate. Moreover, we chose five to test with given that the large majority of our respondents gave five or fewer responses and we wanted to be careful with trendability for a tracking study.

Results

Figure 8: Number of Valid Openends

	U.S.		Japan	
	Mobile	PC	Mobile*	PC
15 Lines	3.9	4.2	**4.0**	3.4
5+ Lines	3.9	4.2	**3.7**	3.3

*p < .05

We were pleased to see that the number of openends was not different by how many lines were shown (Figure 8), and the difference between PC and mobile was not significant in the U.S..

In Japan, we again see no significant differences between number of lines and were surprised to see mobile respondents more verbose than PC respondents. Our reviewer Chris Chapman noted from his own ethnographic research on mobile usage in Japan, that Japanese may be more likely to be engaging on phones in situations allowing longer engagement, such as commuting on trains. Additionally, Japanese may have more familiarity and practice with editing documents on their phones, leading to longer and more survey engagement.

Note that “junk” openends were something we were watching out for and counting, but their incidence was not different by test cell.

Figure 9: Breakoffs

	U.S.		Japan	
	Mobile	PC	Mobile	PC
15 Lines	1.0%	1.2%	**4.0%**	**1.8%**
5+ Lines	2.0%	0.0%	0.6%	1.3%

Our US respondents did not cooperate with our test by abandoning the survey at this question—these rates of drop-off in Figure 9 are actually much lower than what we usually see on this question type. Perhaps we caught them at a better time. Regardless, we see no impact for breakoffs.

We do see in Japan indications that the 5+ lines approach is an improvement here. One other variable we examined was how much time respondents spent on each type and it was approximately the same.

Conclusion

The Japan result led us to feel this was a promising solution for reducing breakoff, and as our results did not change, we went ahead and adopted this approach.

Grid Test

Introduction

Grid questions can be problematic on either a PC or mobile device due to high dropout rates or use of simplification techniques such as straight-lining. Moreover, the optimal approach for PCs may not be the best for mobile devices.

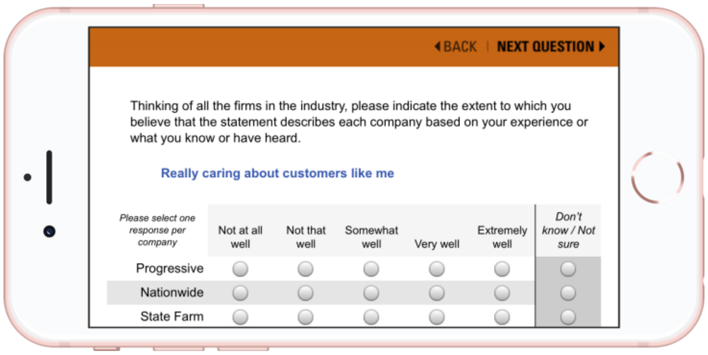
Online interviewing software programs have been advancing new approaches in an attempt to address the presentation of grid questions for mobile surveys.

With the release of Sawtooth Software Lighthouse Studio version 9.1, Lighthouse Studio will detect the size of the respondent's browser window and change the layout of a grid question to become a separate sub question on the same page.

Our grids had been programmed in SSI/Web v8 when the tracker had launched, so we didn't have the benefit of leveraging Lighthouse mobile implementation from the beginning. For the test, Sawtooth Software granted us a trial of Lighthouse Studio.

Our grid test included three cells: grid (control), seen in Figure 10; "scroll" (default Lighthouse behavior) shown in Figure 11; and "freeze" Illustrated in Figure 12. The control grid, programmed as we always had it, would look much like this on a mobile phone, with text extending off the screen whether the phone was held vertically or horizontally.

Figure 10: Grid (control)

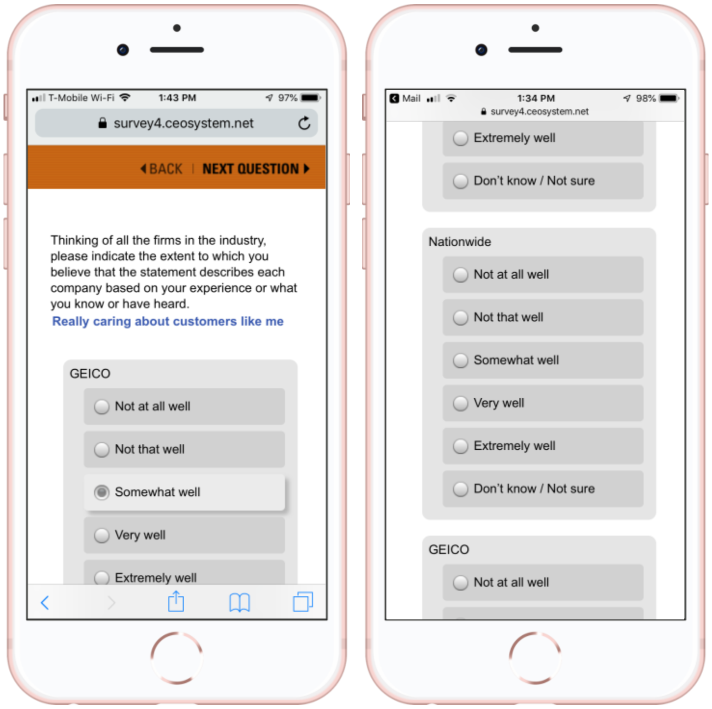


Thinking of all the firms in the industry, please indicate the extent to which you believe that the statement describes each company based on your experience or what you know or have heard.

Really caring about customers like me

Please select one response per company	Not at all well	Not that well	Somewhat well	Very well	Extremely well	*Don't know / Not sure*
Progressive	○	○	○	○	○	○
Nationwide	○	○	○	○	○	○
State Farm	○	○	○	○	○	○

Figure 11: Scroll



Thinking of all the firms in the industry, please indicate the extent to which you believe that the statement describes each company based on your experience or what you know or have heard.

Really caring about customers like me

GEICO

○ Not at all well

○ Not that well

● Somewhat well

○ Very well

○ Extremely well

○ Extremely well

○ Don't know / Not sure

Nationwide

○ Not at all well

○ Not that well

○ Somewhat well

○ Very well

○ Extremely well

○ Don't know / Not sure

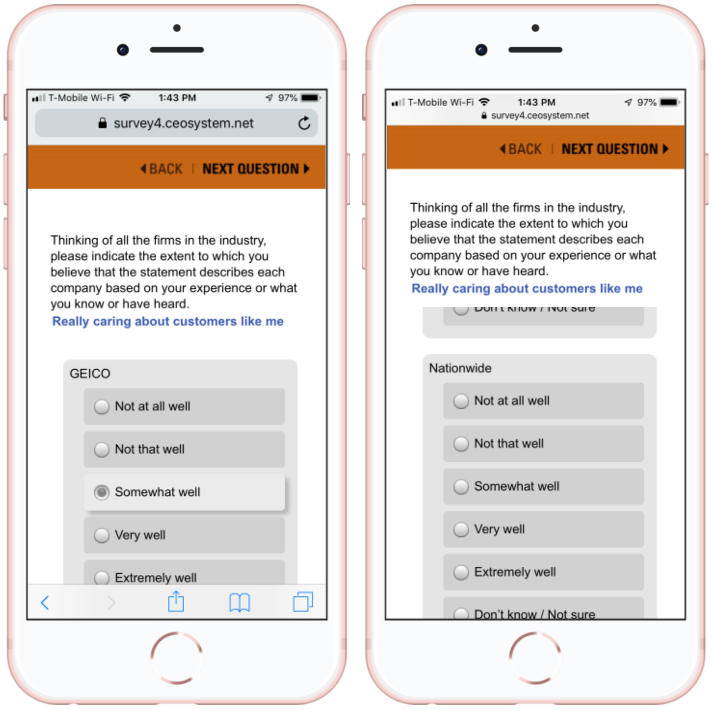
GEICO

○ Not at all well

As we were examining moving to the Lighthouse Studio 9 version of grid for mobile, we also wanted to test using the mobile presentation even for PC. Therefore, each form of the layout was tested for both PC (not shown) and mobile.

However, we had some concerns about how well respondents could answer the question. If you look closely, you can see that as the respondent scrolls, the question and attribute disappear. All that will show is the brand with the response options. You do not see the question nor the attribute under consideration.

Figure 12: Freeze



Thinking of all the firms in the industry, please indicate the extent to which you believe that the statement describes each company based on your experience or what you know or have heard.
Really caring about customers like me

GEICO

○ Not at all well
○ Not that well
● Somewhat well
○ Very well
○ Extremely well

Thinking of all the firms in the industry, please indicate the extent to which you believe that the statement describes each company based on your experience or what you know or have heard.
Really caring about customers like me

○ Don't know / Not sure

Nationwide

○ Not at all well
○ Not that well
○ Somewhat well
○ Very well
○ Extremely well
○ Don't know / Not sure

Therefore, we added a third cell, “freezing” the question and attribute at the top of the screen, even as respondents scrolled down. Thus, what they were answering was always shown.

For this series of attributes, respondents would be rating no more than four brands, with the majority being assigned four based on their familiarity with brands in the insurance industry.

Sample size per cell per device per country ranged from 100 to 113. Respondents were *not* randomly assigned to a device, though they were randomly assigned per condition. Where appropriate, analyses controlled for differences in demographics by device.

Hypotheses

For the grid test we had four hypotheses:

1. Respondents on mobile would be **less likely to break off** with the scrolling approach than with grids.
2. **Straight-lining** behavior would be **reduced** by avoiding grids (though other forms of satisficing might take its place).
3. Consistently displaying the question and attribute (“Freeze”) in the scrolling format would **avoid inconsistency** of mobile responses with responses from grids on PC (compared to “Scroll,” the default).
4. **Scrolling on PC** would not be the best solution for respondent experience but may be **more consistent with mobile scrolling** results. Tradeoff of consistency of results vs. breakoff/respondent experience would then be a point of discussion.

Results

Figure 13: Breakoffs

	U.S.		**Japan**	
	Mobile*	**PC**	**Mobile***	**PC**
Grid	**11%**	3%	**5%**	3%
Scroll	**7%**	1%	**4%**	3%
Freeze	**12%**	4%	**6%**	1%

* p < .05

For breakoffs, both in US and Japan, we saw more break off for mobile than for PC in Figure 13, with no impact of whether it was Grid, Scroll, or Freeze.

Figure 14: Straight-lining

	U.S.		**Japan**	
	Mobile	**PC***	**Mobile**	**PC**
Grid	**8%***	**11%**	8%	4%
Scroll	3%	**5%**	8%	6%
Freeze	3%	**8%**	8%	4%

* p < .05

In the U.S., for straight-lining, as expected, we see high straight-lining for grid and for PC in Figure 14. Mobile respondents were less likely to straight-line while scrolling.

In Japan, there were no effects.

It has been pointed out here that looking across tables, we see the fewest wasted interviews in the U.S. from PC Scroll, with only 1% breaking off and 5% straight-lining.

Figure 15: Consistency in perceptions; most-rated brand top 2 box

Odds	**U.S.**		**Japan**	
Ratio	**Mobile**	**PC**	**Mobile**	**PC**
Grid	**0.6**		**1.4**	
Scroll	**0.5**	**0.6**	**1.4**	**2.5**
Freeze	**0.7**	**0.8**	**1.6**	1.1

In this analysis, with results shown in Figure 15, we are looking at State Farm in the U.S., and Aflac in Japan. These were the brands with which respondents were most familiar. As a default, we are comparing to PC grid, not because it is the right approach but because it was the baseline from our historical tracker, especially from when incidence of mobile starts was much lower.

In the U.S., each of the other cells was significantly less likely to be top two box than the grid question. Mobile Scroll and PC Scroll were nicely aligned (tested independently), as were Mobile Freeze and PC Freeze, and we might have jumped to certain conclusions here; ***but*** in Japan we see the opposite, where most of the other designs yielded a higher likelihood of a top 2 box rating than PC Grid, and Mobile Scroll and PC Scroll are *not* aligned.

Note that for the consistency analysis, to isolate the effect of the design conditions, estimated models controlled for gender, age, household income, employment status for both, and additional for US race and region of the country.

In addition, we asked two questions at the end of the survey, following the example of a paper presented at AAPOR (Sarraf, Brooks, Cole, & Wang, 2015):

"To quantify what you just wrote, considering factors like ease of navigation, ease of reading the screen, and ease of selecting responses, please rate how easy it was for you to complete this survey."

There were no differences on this question for either geography.

Figure 16: Visual Design (4-point scale)

	U.S.		Japan	
Mean	**Mobile**	**PC**	**Mobile**	**PC**
Grid*	**3.3**	**3.4**	2.7	2.8
Scroll	3.3	3.1	2.6	2.7
Freeze	3.1	3.2	2.6	2.7

* p < .05

The second question we asked, on a 4-point scale, was: "And how would you rate the visual design of this survey?"

In the U.S. in Figure 16 we see an effect for layout, with Grid rated significantly higher than Scroll or Freeze. There were no differences in Japan.

Considerations

We do have other things to consider, however.

Scrolling added a moderate amount of time to the survey, plus 1-2 minutes beyond an otherwise 14-minute survey and freezing the top of the question added an additional 2-3 minutes.

Additionally, we gave the respondents an opportunity to give their feedback and they were happy to do so. Those who took the mobile-optimized layouts on PC expressed a preference for seeing all the questions and answers together in one place.

Those on their phones were fairly positive overall about the experience, but noted various adaptations we might consider, or that they needed to undertake to easily take the survey.

Conclusions

To summarize our findings,

1. We did not find respondents less likely to break off when scrolling than for grid.
2. We *did* find straight-lining to be reduced when grid items were replaced by scrolling sets of single-response items, but only in the U.S..
3. In looking at consistency with PC Grid responses, we examined how often respondents chose a top two box response. Respondents receiving the Freeze design give more similar

responses to PC Grid than those who received the Scroll design do but are still significantly less likely to select top two box in the U.S. and significantly *more* likely to select top two box in Japan. On the other hand, both Freeze and Scroll are more consistent across device than the grid format is, particularly in the U.S.

4. Scrolling on PC was consistent in the U.S. with scrolling on mobile, but not in Japan. However, the additional time to take the survey and the respondent feedback requesting grids make us unlikely to adopt the scrolling approach.

For grid, we concluded to maintain our current (grid on mobile) approach and keep looking as:

- The "Freeze" approach was untenable due to negative respondent quantitative and qualitative feedback as well as the extra time it took to complete the survey, and
- The "Scroll" design on mobile had one advantage in reducing straight lining, but otherwise did not yield a meaningful advantage, and took longer. Moreover, for Japan, the data was wildly different than other forms.

It should be noted that one size does not fit all, as one set of authors (Mavletova, Couper, & Lebedev, 2018) noted: ". . . there are a large number of factors that may affect the choice of a grid format for surveys, including the length and complexity of the survey, the number of items, the number of response options, the proportion of the respondents using a mobile device, the type of mobile optimization (if any) for smartphones, and so on. The choice of whether or not to use grid question should be made on a case-by-case basis and is not an all-or-nothing decision. If the survey software does not optimize for mobile devices (particularly smartphones), we suggest that using an item-by-item format for both mobile and PC web may result in lower measurement error and higher measurement equivalence in a survey."

Practical Tips for Success

1. **Know your audience**. Will they want to use smartphones to complete the survey? Will you want those who prefer phones in your data?
2. Whether you plan to collect data using phones or not, if a panel audience, **have a conversation with your provider** during kickoff regarding how data will be collected.
3. **Test** the survey on *various* mobile devices, browsers, and operating systems. Common oversights include:
    a. Respondent instructions that don't apply for mobile, e.g., "click here," "use your mouse to . . . ," or, "select one in each row" when question has been converted to a series of single response questions.
    b. Grids that may wrap response points.
    c. Referencing non-existing point labels (e.g., 5=strongly agree but in the mobile version, response options are unnumbered).

For Further Research

1. On the grid test, we could have considered freezing *only* the attribute and not the whole question and header. This would have reduced the amount of scrolling the respondent had to do while keeping the most relevant information on the screen.

2. The respondent experience and survey results may differ depending how many brands respondents are asked to rate.
3. Other grid types are available to test, using different software:
    a. **Accordion grids** (response options auto-expand and collapse item by item as the respondent clicks).
    b. **Progressive grids** (attributes auto-advance while response options remain stationary).
4. Or don't use grids at all, instead use **MaxDiff/Best-Worst**. SKIM (Ruitenburg, Joost van (SKIM), 2018) has developed additional options for mobile.

Deb Ploskonka

Karlan Witt

BIBLIOGRAPHY

Kantar IMRB. (2019). *21st edition ICUBE(TM) Digital adoption & usage trends.* Retrieved from https://imrbint.com/images/common/ICUBE%E2%84%A2_2019_Highlights.pdf

Mavletova, A., Couper, M. P., & Lebedev, D. (2018). Grid and Item-by-Item Formats in PC and Mobile Web Surveys. *Social Science Computer Review, 36*(6), 647–668.

McKinsey Global Institute. (2019, March). *Digital India: Technology to transform a connected nation.* McKinsey & Company.

Peterson, G., Griffin, J., LaFrance, J., & Li, J. (2017). Chapter 10: Smartphone Participation in Web Surveys: Choosing Between the Potential for Coverage, Nonresponse, and Measurement Error. In P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, . . . B. T. West, *Total Survey Error in Practice* (pp. 203–233). Hoboken, New Jersey: John Wiley & Sons, Inc.

Pew Research Center. (2019, June 12). *Internet/Broadband Fact Sheet*. Retrieved from Pew Research: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/

Pew Research Center. (2019, June 13). *Mobile Technology and Home Broadband 2019*. Retrieved from Pew Research Center: https://www.pewinternet.org/2019/06/13/mobile-technology-and-home-broadband-2019/

Poynter, R., Williams, N., & York, S. (2014). The Handbook of Mobile Market Research: Tools and Techniques for Market Researchers. Chichester: Wiley.

Ruitenburg, Joost van (SKIM). (2018, April 26). Optimizing Conjoint for Mobile: Mixed Profile Swiping [Webinar]. SKIM. Retrieved from https://skimgroup.com/events/webinar-introducing-a-better-mobile-first-approach-for-conjoint-research/

Sarraf, S., Brooks, J., Cole, J., & Wang, X. (2015, May 16). What is the impact of smartphone optimization on long surveys? *Presentation to the American Association for Public Opinion Research Annual Conference*. Hollywood, FL.

A RESEARCHER'S GUIDE TO STUDYING LARGE ATTRIBUTE SETS IN CHOICE-BASED CONJOINT

MEGAN PEITZ
NUMERIOUS
MIKE SERPETTI
DAN YARDLEY
GONGOS

ABSTRACT

There is no arguing that choice-based methods have become dominant in our industry. Yet, there is no clear answer on what a researcher should do as the number of attributes increases (>6). Design techniques including Partial Profile and Adaptive Choice-Based Conjoint offer solutions, but past research has yet to crown a winner. This paper sets out to revisit prior research and to explore and validate both Partial Profile (PP) and Adaptive Choice-Based Conjoint (ACBC) in comparison to Full Profile (FP) Choice-Based Conjoint (CBC) with real respondents on experiments of 10, 15, and 20 attributes to determine which method is best across multiple scenarios.

INTRODUCTION

Most of the choice-based research done today is Full Profile (FP), where a level from every attribute is shown in every product profile. However, some argue that there comes a point when a FP choice task is too cumbersome and overwhelming, forcing respondents to use a simplification heuristic that could affect the model's predictability. Since the work of Green and Srinivasan (Green, P. & Srinivasan, 1978), we have been historically taught to use around six attributes (depending on level text, category, and more). Two solutions to this problem include Partial Profile (PP) and Adaptive Choice-Based Conjoint (ACBC). PP is where a level from only a subset of attributes, usually 7 or fewer, is shown in every product profile. The subset of attributes changes across every screen so that respondents evaluate all attributes, but only 7 at a time (Chrzan, K. & Elrod, T., 1995).

ACBC takes respondents through three main phases: 1) BYO (configuration) phase, 2) Consideration phase, and 3) Choice Tournament phase, adapting the design depending upon answers within these phases to account for non-compensatory decision making that can happen (Johnson, R. & Orme, B., 2007). Screen shots of the FP CBC, PP CBC, and ACBC exercises can be found in the appendix.

STUDY DESIGN

We chose smartphones as the product and created sample cells with 10 attributes, 15 attributes, and 20 attributes to describe the product profiles. The set of attributes and levels tested is in Figure 1.1.

Figure 1.1: Attributes and Levels Tested

10	15	20	ATTRIBUTE	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5	LEVEL 6
X	X	X	**Brand**	Apple	Samsung	Motorola	Google	LG	Sony
X	X	X	**Price**	$700	$800	$900	$1,000	$1,100	
X	X	X	**Screen Size**	4.6"	5.2"	5.8"	6.4"		
X	X	X	**Storage**	64 GB	128 GB	256 GB	512 GB	1TB	
X	X	X	**Quality of Camera**	8 megapixels	12 megapixels	16 megapixels	24 megapixels		
X	X	X	**Generation**	4G	5G				
X	X	X	**Battery Life**	14 hours	20 hours	26 hours	32 hours	40 hours	
X	X	X	**RAM**	2GB RAM	4GB RAM	6GB RAM	8GB RAM		
X	X	X	**HDR**	No	Yes				
X	X	X	**Waterproof**	No	Yes				
	X	X	**Wireless Charging**	No	Yes				
	X	X	**Headphone Jack**	No	Yes				
	X	X	**Screen Quality**	Standard (800x400)	High Def (1280x720)	Full HD (1920x1080)			
	X	X	**Dual Camera**	No	Yes				
	X	X	**SD Slot**	No	Yes				
		X	**Color**	White	Black	Blue	Gold	Pink	Silver
		X	**Weight**	5 oz	6 oz	7 oz	8 oz		
		X	**Facial Recognition**	Touch ID	Face ID				
		X	**Display**	LCD	OLED				
		X	**Video Recording Quality**	Up to 720p	Up to 1080p	Up to 4k			

Each cell was assigned a specific methodology, shown in Figure 1.2.

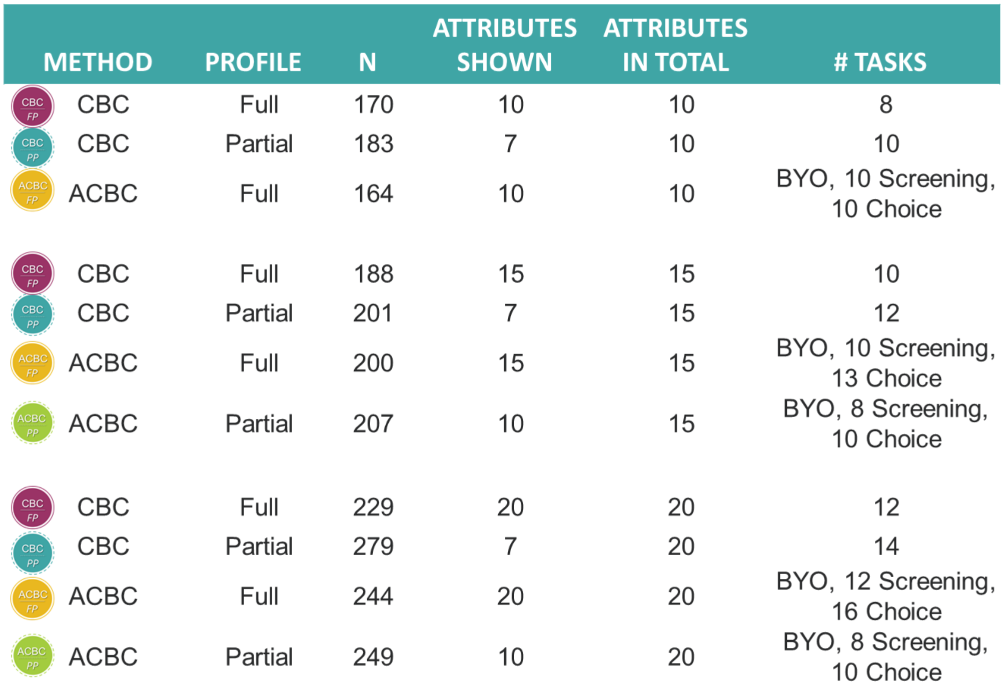
Figure 1.2: Methodologies Tested

CBC *FP*	CHOICE-BASED CONJOINT *FULL PROFILE*	All attributes and all levels are shown in every product profile.
CBC *PP*	CHOICE-BASED CONJOINT *PARTIAL PROFILE*	The computer randomly selects a subset of 7 attributes to show in every product profile. Each screen shows a different subset; Brand & Price were not forced.
ACBC *FP*	ADAPTIVE CHOICE-BASED CONJOINT *FULL PROFILE*	All attributes and all levels are shown in every product profile.
ACBC *PP*	ADAPTIVE CHOICE-BASED CONJOINT *PARTIAL PROFILE*	Each respondent chooses a subset of attributes that are important to them, between 2 and 10. Only those attributes are shown in every product profile. Each screen shows the same subset; Brand & Price were not forced.

Sample Cells

There were 11 sample cells in total. Three with 10 attributes, four with 15 attributes, and four with 20 attributes. See Figure 1.3 for detail on the sample cells. Each sample cell was weighted equally by gender, age, and ethnicity. Sample was provided by Dynata.

Figure 1.3: Sample Cell Overview



	METHOD	PROFILE	N	ATTRIBUTES SHOWN	ATTRIBUTES IN TOTAL	# TASKS
CBC FP	CBC	Full	170	10	10	8
CBC PP	CBC	Partial	183	7	10	10
ACBC FP	ACBC	Full	164	10	10	BYO, 10 Screening, 10 Choice
CBC FP	CBC	Full	188	15	15	10
CBC PP	CBC	Partial	201	7	15	12
ACBC FP	ACBC	Full	200	15	15	BYO, 10 Screening, 13 Choice
ACBC PP	ACBC	Partial	207	10	15	BYO, 8 Screening, 10 Choice
CBC FP	CBC	Full	229	20	20	12
CBC PP	CBC	Partial	279	7	20	14
ACBC FP	ACBC	Full	244	20	20	BYO, 12 Screening, 16 Choice
ACBC PP	ACBC	Partial	249	10	20	BYO, 8 Screening, 10 Choice

Design Strategy

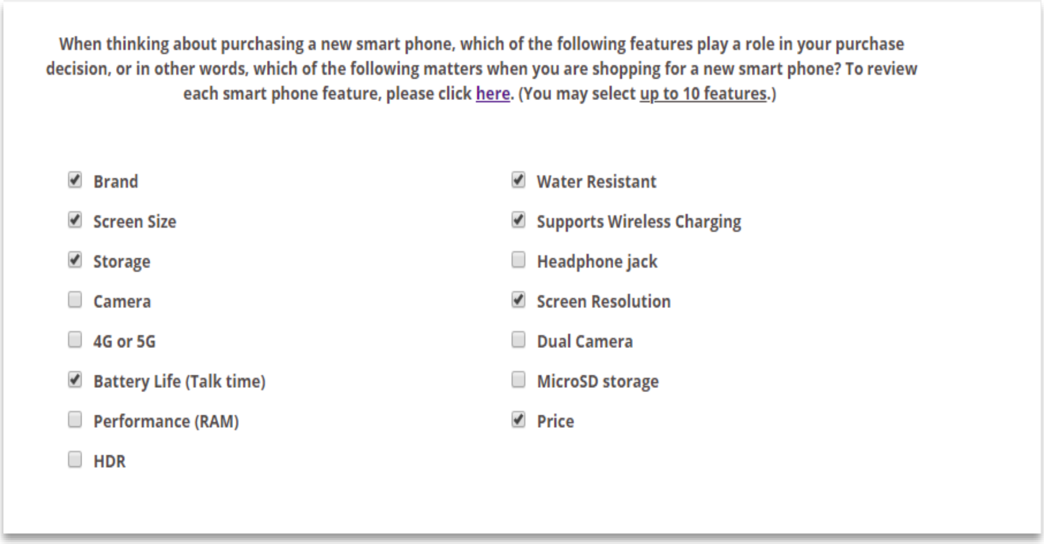
The FP CBC design was generated using Sawtooth Software's Shortcut design algorithm. The PP CBC design was generated using Sawtooth Software's Complete Enumeration design algorithm (as the attributes not shown in the product profile are essentially assuming level overlap). The ACBC designs were generated using Sawtooth Software's default settings.

Partial Profile Attribute Selection

For the PP CBC, the computer selected 7 attributes on each screen based upon the design algorithm. Brand and Price were not forced onto every screen in the PP CBC.

For the PP ACBC, respondents could choose between 2 to 10 attributes. Brand and Price were not forced into the respondent's subset for PP ACBC. Figure 1.4 shows an example of how respondents configured their subset.

Figure 1.4: PP ACBC 15 Attribute Selection Question



When thinking about purchasing a new smart phone, which of the following features play a role in your purchase decision, or in other words, which of the following matters when you are shopping for a new smart phone? To review each smart phone feature, please click here. (You may select up to 10 features.)

☑ Brand
☑ Screen Size
☑ Storage
☐ Camera
☐ 4G or 5G
☑ Battery Life (Talk time)
☐ Performance (RAM)
☐ HDR

☑ Water Resistant
☑ Supports Wireless Charging
☐ Headphone jack
☑ Screen Resolution
☐ Dual Camera
☐ MicroSD storage
☑ Price

15 Attribute PP ACBC Selection

Respondents vary in the number of attributes they find play a role in their decision to purchase a smartphone. 71% include Price and only 60% include Brand. Figure 1.5 shows the distribution of the number of attributes per respondent. Figure 1.6 shows the proportion of respondents that chose that attribute.

Figure 1.5: % of # of Attributes in 15 Attribute PP ACBC per Respondent

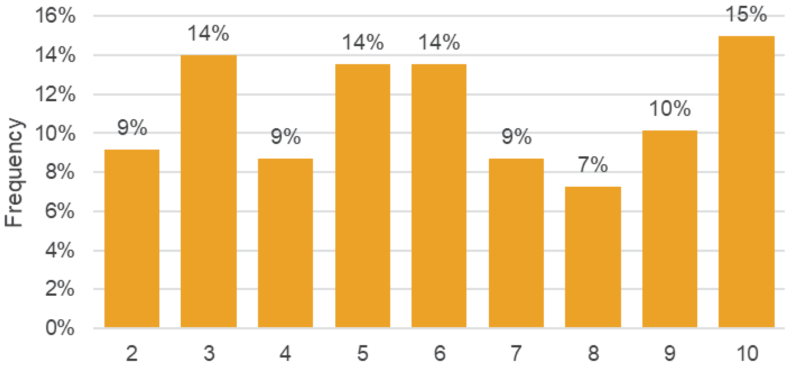
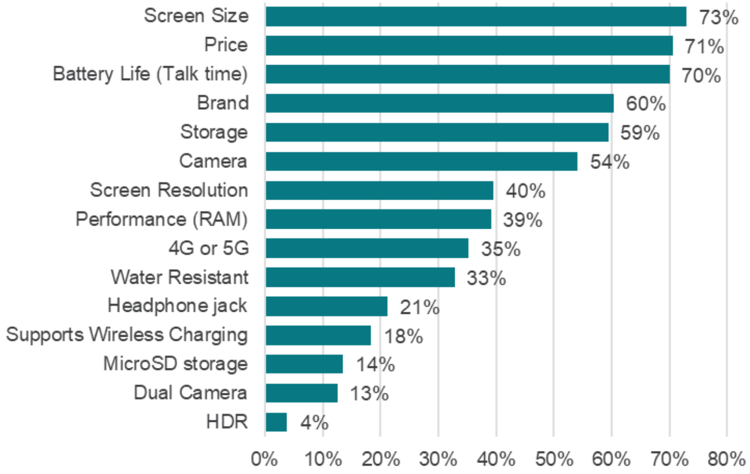


Figure 1.6: % of Respondents Including Each Attribute in Their Subset



20 Attribute PP ACBC Selection

When shown 20 attributes, respondents are more likely to choose the maximum number of attributes allowed (10) in comparison to 15 attributes. 66% include Price and 61% include Brand. Figure 1.7 shows the distribution of the number of attributes per respondent. Figure 1.8 shows the proportion of respondents that chose that attribute.

Figure 1.7: % of # of Attributes in 20 Attribute PP ACBC per Respondent

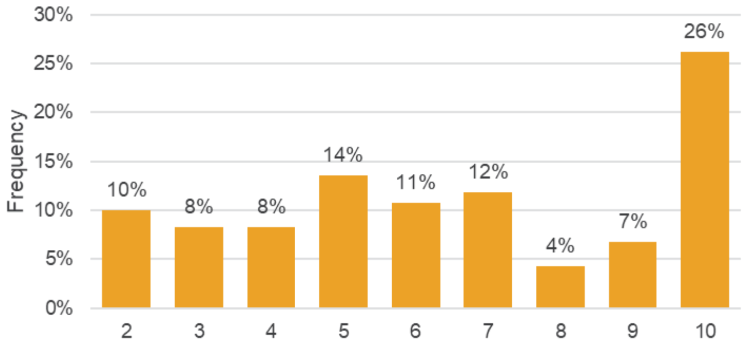
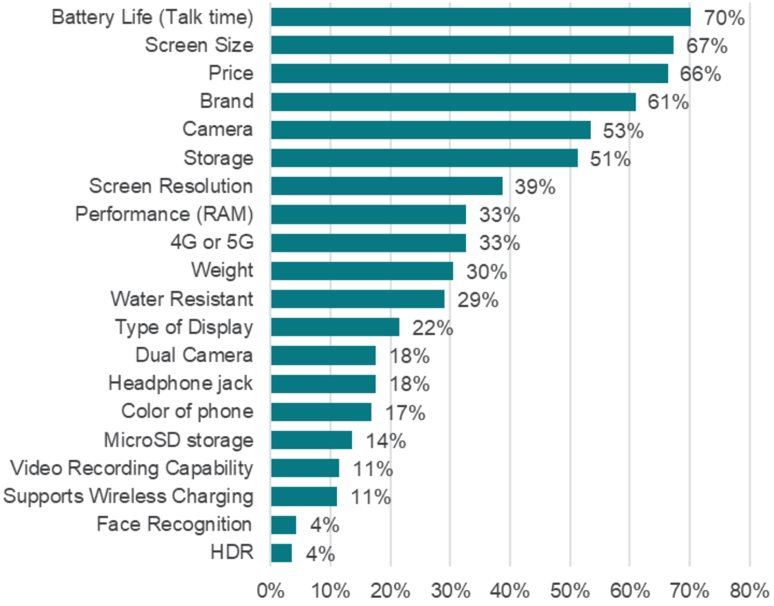


Figure 1.8: % of Respondents Including Each Attribute in Their Subset

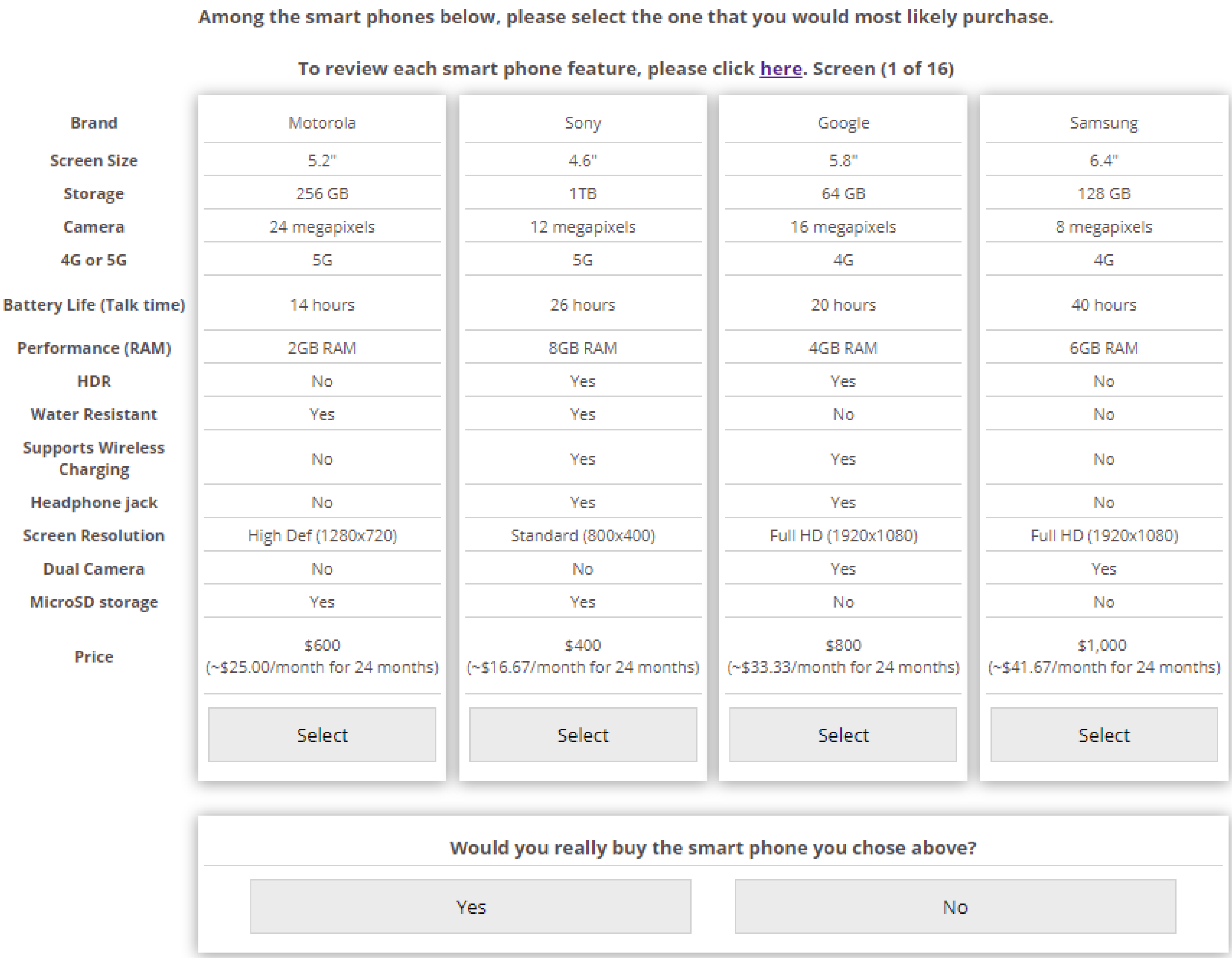


In future projects, analysts should consider forcing brand and price into every subset of the PP CBC and list of attributes in the PP ACBC. Doing this may depict a more accurate representation of the real world, given the influence these attributes have on decision making.

HOLDOUT TASKS

Holdout tasks can be used to compare the predictive validity of one conjoint method to another. In a holdout task, the researcher specifies exactly which combinations of attribute levels to show in a product profile and all respondents see this exact scenario. Chrzan (2015) suggests that at least 5 holdout tasks, if not more, are needed to be confident in these conclusions. This study includes 6 CBC-looking holdout tasks (Figure 2.1).

Figure 2.1: CBC Holdout Example



Among the smart phones below, please select the one that you would most likely purchase.

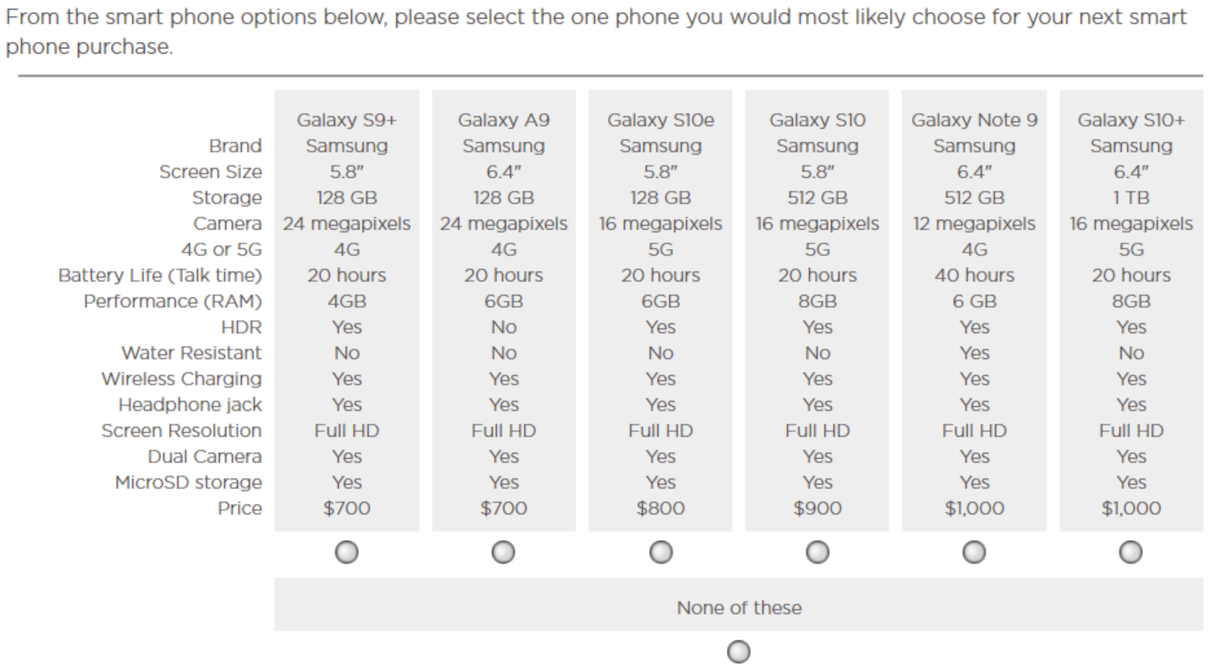
To review each smart phone feature, please click here. Screen (1 of 16)

Brand	Motorola	Sony	Google	Samsung
Screen Size	5.2"	4.6"	5.8"	6.4"
Storage	256 GB	1TB	64 GB	128 GB
Camera	24 megapixels	12 megapixels	16 megapixels	8 megapixels
4G or 5G	5G	5G	4G	4G
Battery Life (Talk time)	14 hours	26 hours	20 hours	40 hours
Performance (RAM)	2GB RAM	8GB RAM	4GB RAM	6GB RAM
HDR	No	Yes	Yes	No
Water Resistant	Yes	Yes	No	No
Supports Wireless Charging	No	Yes	Yes	No
Headphone jack	No	Yes	Yes	No
Screen Resolution	High Def (1280x720)	Standard (800x400)	Full HD (1920x1080)	Full HD (1920x1080)
Dual Camera	No	No	Yes	Yes
MicroSD storage	Yes	Yes	No	No
Price	$600 (~$25.00/month for 24 months)	$400 (~$16.67/month for 24 months)	$800 (~$33.33/month for 24 months)	$1,000 (~$41.67/month for 24 months)

Would you really buy the smart phone you chose above?

Since the best measure of success is each model's ability to predict real-world market shares, two shelf holdouts were also included to reflect the actual consumer decision-making process. The first shelf is a Full Profile and includes 20 total products throughout the duration of the exercise (Figure 2.2 shows one of these screens that tests 6 Samsung Galaxy products).

Figure 2.2: FP Shelf-Holdout Example



From the smart phone options below, please select the one phone you would most likely choose for your next smart phone purchase.

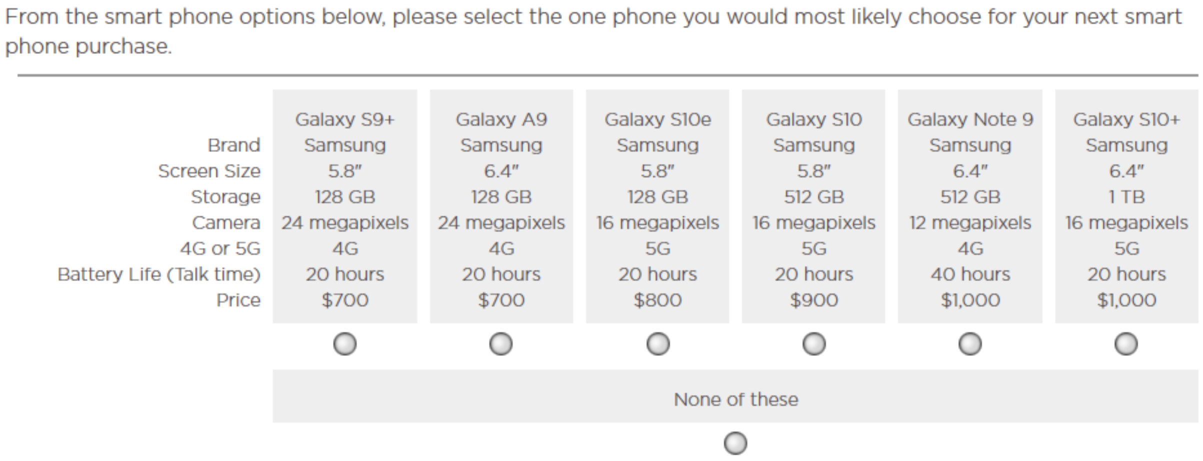
	Galaxy S9+	Galaxy A9	Galaxy S10e	Galaxy S10	Galaxy Note 9	Galaxy S10+
Brand	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung
Screen Size	5.8″	6.4″	5.8″	5.8″	6.4″	6.4″
Storage	128 GB	128 GB	128 GB	512 GB	512 GB	1 TB
Camera	24 megapixels	24 megapixels	16 megapixels	16 megapixels	12 megapixels	16 megapixels
4G or 5G	4G	4G	5G	5G	4G	5G
Battery Life (Talk time)	20 hours	20 hours	20 hours	20 hours	40 hours	20 hours
Performance (RAM)	4GB	6GB	6GB	8GB	6 GB	8GB
HDR	Yes	No	Yes	Yes	Yes	Yes
Water Resistant	No	No	No	No	Yes	No
Wireless Charging	Yes	Yes	Yes	Yes	Yes	Yes
Headphone jack	Yes	Yes	Yes	Yes	Yes	Yes
Screen Resolution	Full HD	Full HD	Full HD	Full HD	Full HD	Full HD
Dual Camera	Yes	Yes	Yes	Yes	Yes	Yes
MicroSD storage	Yes	Yes	Yes	Yes	Yes	Yes
Price	$700	$700	$800	$900	$1,000	$1,000
	○	○	○	○	○	○

None of these

○

Research has shown that FP is better than PP at predicting hit rates for FP holdout choice tasks, therefore a PP shelf was also included so to not tip the hat in favor of FP. In the PP shelf, each product profile is defined by the same 7 attributes across all respondents and includes 20 total products throughout the duration of the exercise (Figure 2.3 shows one of these screens that tests 6 Samsung Galaxy products).

Figure 2.3: PP Shelf-Holdout Example



From the smart phone options below, please select the one phone you would most likely choose for your next smart phone purchase.

	Galaxy S9+	Galaxy A9	Galaxy S10e	Galaxy S10	Galaxy Note 9	Galaxy S10+
Brand	Samsung	Samsung	Samsung	Samsung	Samsung	Samsung
Screen Size	5.8″	6.4″	5.8″	5.8″	6.4″	6.4″
Storage	128 GB	128 GB	128 GB	512 GB	512 GB	1 TB
Camera	24 megapixels	24 megapixels	16 megapixels	16 megapixels	12 megapixels	16 megapixels
4G or 5G	4G	4G	5G	5G	4G	5G
Battery Life (Talk time)	20 hours	20 hours	20 hours	20 hours	40 hours	20 hours
Price	$700	$700	$800	$900	$1,000	$1,000
	○	○	○	○	○	○

None of these

○

In addition, we must remember that because each of these tasks look like CBC tasks, there is a potential methods effect in favor of CBC vs. ACBC.

The Models

In all 11 cells, we created a hierarchical Bayesian (HB) model with prior variance of 0.5 and 5 degrees of freedom and used point estimates (the default Sawtooth Software settings). In addition, no prohibitions, constraints, or interactions were included in any model.

For the scale factor (response error) involved in the different calibration tasks, and holdout task layouts 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

Comparisons across the methodologies are made within the cells that test 10 attributes, 15 attributes, and 20 attributes. Statistically, the most important comparison is how well the models perform. Here, we compare the Mean Absolute Error (MAE) and how well the model predicts the None category. In addition, price sensitivity curves and Willingness-to-Pay values are compared.

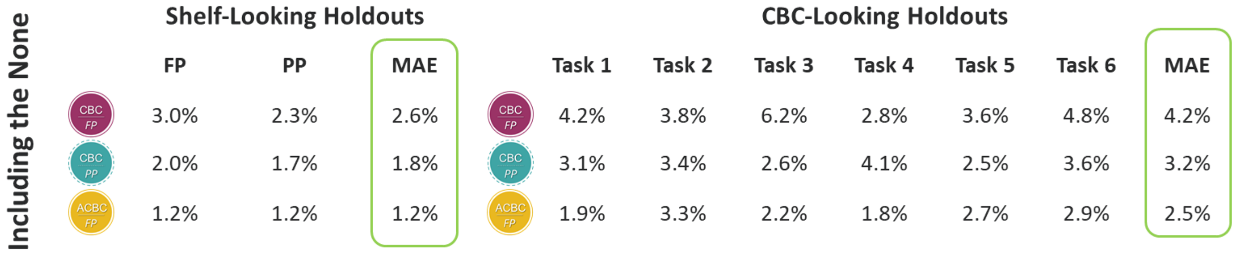
However, it is also just as important to have an enjoyable respondent experience. Therefore, we will also compare the methodologies from the respondent's perspective, examining median time to complete/length of interview (LOI), drop-off percentages, percentage of those who admit to cheating, and respondent evaluations (i.e., easy vs. hard, fun vs. dull).

The 10 Attribute Story

Model Validity

When simulating the shelf and CBC holdout tasks and comparing the simulated shares to the actual holdout shares, we find that ACBC has the lowest MAE (Table 3.1). These simulations include the None parameter (i.e., respondents can opt out of buying a phone).

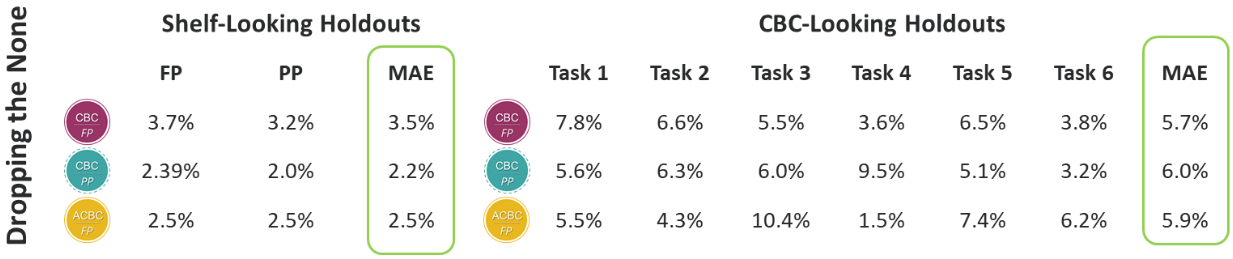
Table 3.1 Mean Absolute Error Across 10 Attribute Shelf and CBC Holdouts Including the None



		Shelf-Looking Holdouts				CBC-Looking Holdouts						
		FP	PP	MAE		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	MAE
Including the None	CBC FP	3.0%	2.3%	2.6%	CBC FP	4.2%	3.8%	6.2%	2.8%	3.6%	4.8%	4.2%
	CBC PP	2.0%	1.7%	1.8%	CBC PP	3.1%	3.4%	2.6%	4.1%	2.5%	3.6%	3.2%
	ACBC FP	1.2%	1.2%	1.2%	ACBC FP	1.9%	3.3%	2.2%	1.8%	2.7%	2.9%	2.5%

However, because the None parameter is computed differently within ACBC vs. CBC (i.e., in ACBC the None is determined from the number of concepts marked as a possibility vs. not a possibility in the screening section), we sought to compare the MAEs when excluding the None option in simulations. After dropping the None, the MAEs for all three methods are comparable (Table 3.2). Again, we should note that shared methods bias would be expected to favor FP and PP CBC rather than ACBC, because the holdout tasks involved CBC-looking tasks.

Table 3.2 Mean Absolute Error Across 10 Attribute Shelf and CBC Holdouts Excluding the None

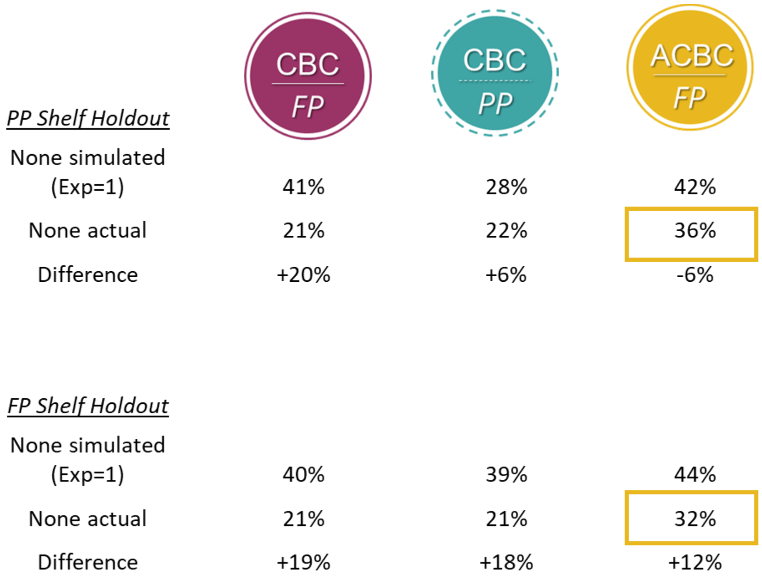


		Shelf-Looking Holdouts				CBC-Looking Holdouts						
		FP	PP	MAE		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	MAE
Dropping the None	CBC FP	3.7%	3.2%	3.5%	CBC FP	7.8%	6.6%	5.5%	3.6%	6.5%	3.8%	5.7%
	CBC PP	2.39%	2.0%	2.2%	CBC PP	5.6%	6.3%	6.0%	9.5%	5.1%	3.2%	6.0%
	ACBC FP	2.5%	2.5%	2.5%	ACBC FP	5.5%	4.3%	10.4%	1.5%	7.4%	6.2%	5.9%

The None Parameter

Taking a closer look at the None, we find that those respondents in the ACBC sample cells are significantly more likely to choose the None (i.e., respondents would not buy any product on the shelf) (Table 3.3). This could signal a psychological effect of the ACBC exercise, making them less likely to buy an actual product after building their own. Further research should be done to determine if this is category specific, or methodology specific.

Table 3.3: None Proportions within the 10 Attribute Experiments



	CBC FP	CBC PP	ACBC FP
PP Shelf Holdout			
None simulated (Exp=1)	41%	28%	42%
None actual	21%	22%	36%
Difference	+20%	+6%	-6%
FP Shelf Holdout			
None simulated (Exp=1)	40%	39%	44%
None actual	21%	21%	32%
Difference	+19%	+18%	+12%

Price Sensitivity

Price sensitivity is a primary deliverable of choice research. Therefore, we simulated one product versus the None and graphed the share of preference estimated from each model when only changing price.

Figure 3.4: Price Sensitivity of 10 Attributes by Methodology

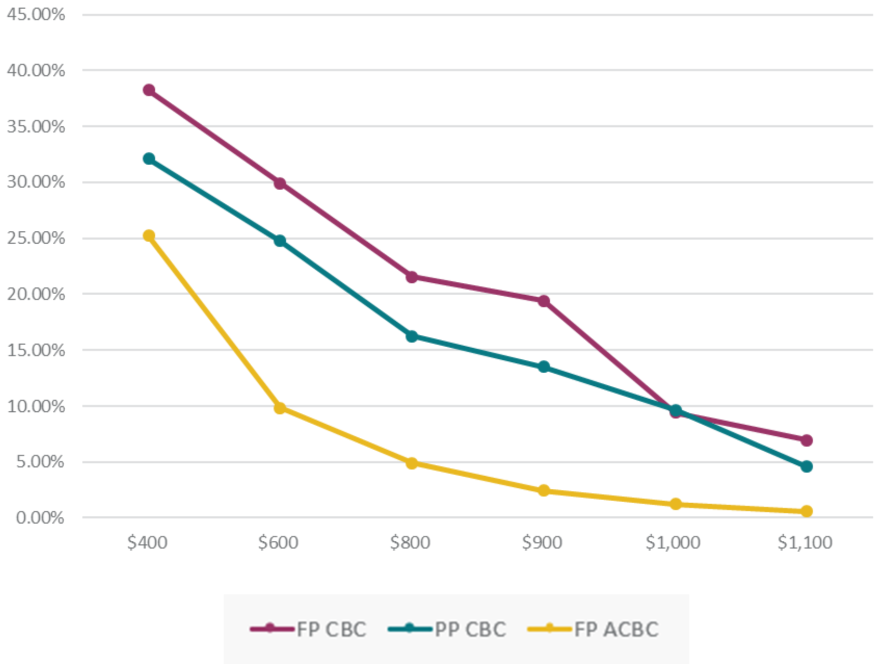


Figure 3.4 shows that FP ACBC seems to conservatively predict choice compared to FP CBC and PP CBC. This is in alignment with the high None proportion found in the ACBC data. FP CBC and PP CBC seem to have parallel curves—even though price was not forced into every product profile for PP CBC.

Willingness to Pay (WTP)

Although there are many caveats to calculating WTP data, the authors wanted to explore any differences in the methodologies tested. Our WTP data is calculated using the simulation approach where two products with the exact same specs are simulated versus the None. By default, their shares will be exactly equal. We then add a feature to the second product that the first product doesn't have, (i.e., increase product 2 from 4G to 5G; product 1 stays at 4G) and find the price for the second product at which shares for both products return to equal.

Figure 3.5: WTP for Brand, Storage, and Generation Attributes by Methodology

	CBC FP	CBC PP	ACBC FP
BRAND			
APPLE	$200	$200	$200
SAMSUNG	$55	$109	$200
GOOGLE	$0	$0	$0
LG	-$181	$11	-$67
MOTOROLA	-$262	-$310	-$130
SONY	-$322	-$238	-$90
STORAGE			
64 GB	-$172	-$222	-$28
128 GB	$0	$0	$0
256 GB	$32	$7	$33
512 GB	$41	$29	$138
1TB	$47	-$11	$117
GENERATION			
4G	$0	$0	$0
5G	$16	$49	$63

The results show that FP ACBC has a WTP that is flatter than both FP CBC and PP CBC. PP CBC seems to align well with FP CBC. This is also found when comparing correlations between utilities for the different methods (Figure 3.6).

Figure 3.6: Correlations of Aggregate Utilities by Methodology

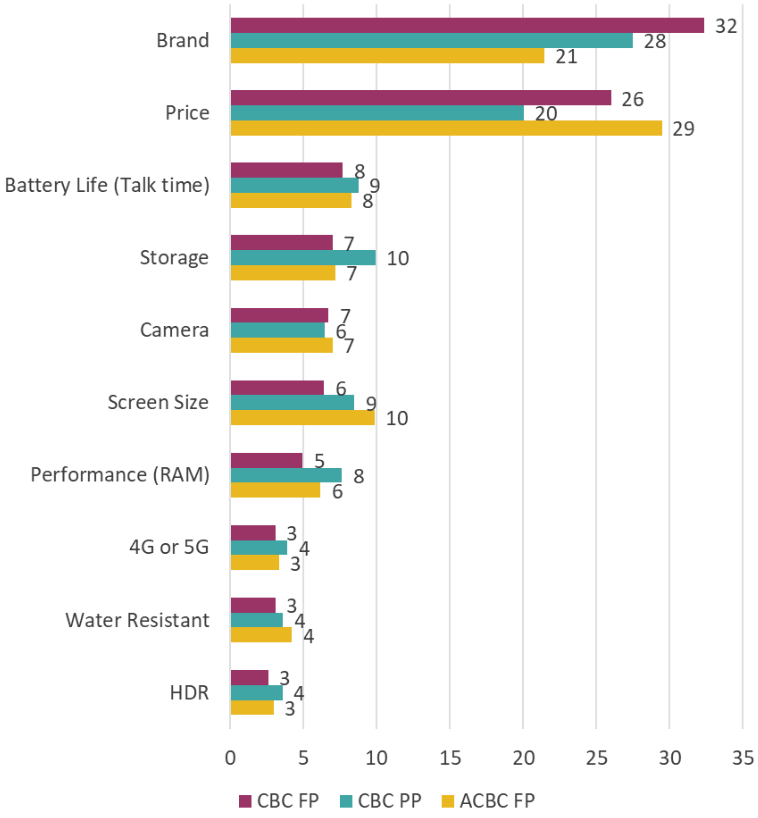
Utility Correlations	CBC FP	CBC PP	ACBC FP
CBC FP	1		
CBC PP	0.94	1	
ACBC FP	0.89	0.92	1

Importance Scores

Importance scores are a typical deliverable, albeit with many potential issues (i.e., lies are in the averages, extremely influenced by the levels tested). While the authors do not believe this is the best way to look at the data, it is interesting to see the differences since all attributes/levels tested across the methods are the same, the importance scores shown are the standard Sawtooth Software calculation of importance scores (i.e., range of HB utilities per attribute per respondent). The authors realize that there are different ways to calculate attribute importance, but wanted to investigate the results based on how a typical conjoint user would use attribute importance.

As expected, FP CBC has the steepest importance scores. Interestingly, FP CBC and PP CBC find Brand as the most important attribute followed by Price, where FP ACBC finds Price more important than Brand (Figure 3.7).

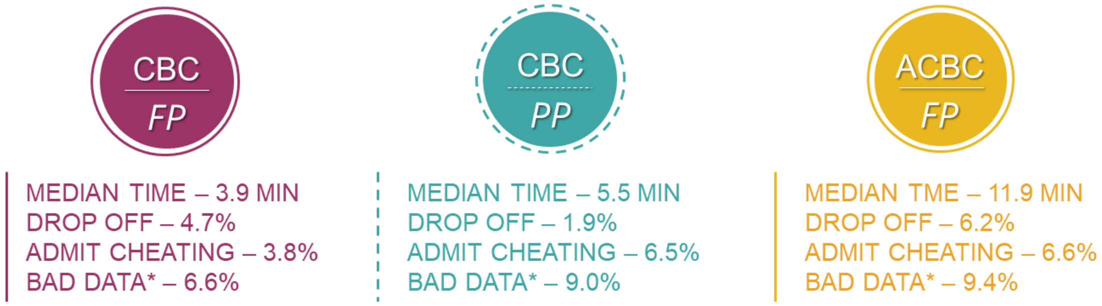
Figure 3.7: Importance Scores by Methodology



Respondent Preference

FP CBC is the quickest exercise to complete, while FP ACBC is the longest (Figure 3.8). FP ACBC also has the highest drop-off percentage, highest likelihood of cheating, and the highest proportion of bad data. 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.

Figure 3.8: Respondent Statistics for 10 Attribute Experiments



MEDIAN TIME – 3.9 MIN
DROP OFF – 4.7%
ADMIT CHEATING – 3.8%
BAD DATA* – 6.6%

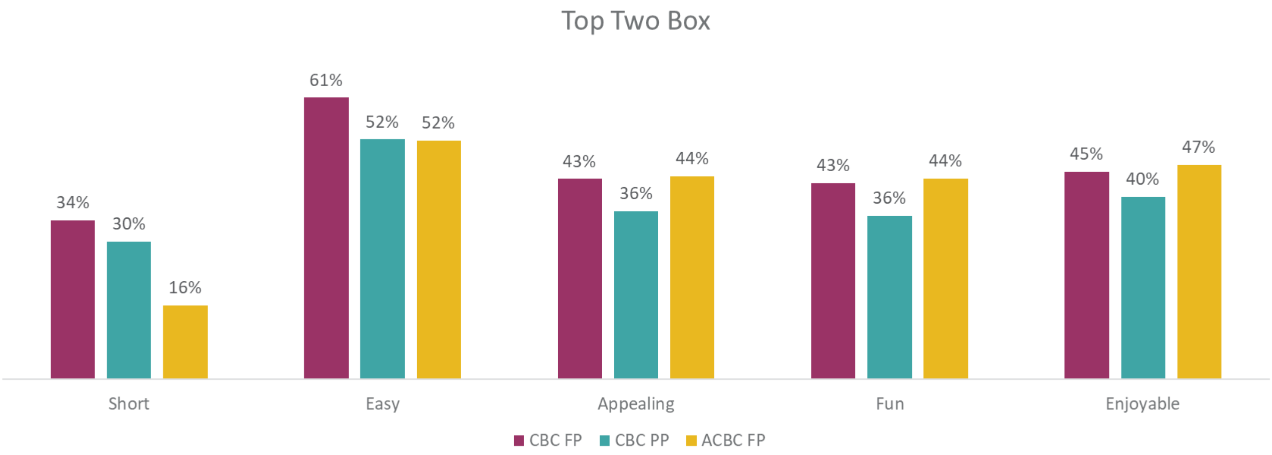
MEDIAN TIME – 5.5 MIN
DROP OFF – 1.9%
ADMIT CHEATING – 6.5%
BAD DATA* – 9.0%

MEDIAN TME – 11.9 MIN
DROP OFF – 6.2%
ADMIT CHEATING – 6.6%
BAD DATA* – 9.4%

Missed 2 data quality checks or admitted cheating

When rating the different methodologies, FP CBC is the shortest and easiest, while FP ACBC relatively more Enjoyable, Fun, and Appealing (Figure 3.9). Both FP CBC and FP ACBC seem to edge out PP CBC, but the differences are not significant.

Figure 3.9: Respondent Top Two Agreement on 10 Attribute Survey Experience



10 Attribute Conclusion

Overall, FP CBC seems to perform slightly better when testing 10 attributes, from both a respondent and model perspective. (Again, with the caveat that the holdouts would be expected to favor the CBC-looking approaches and put ACBC at a disadvantage.)

15 Attribute Story

Model Validity

Table 4.1 shows that when including the None option in simulations, ACBC has the lowest MAE, particularly PP ACBC. After dropping the None, the MAE for the Adaptive methods are more in line with FP CBC (Table 4.2). PP CBC has the highest MAE, perhaps due to the number of attributes shown out of the total attributes modeled (7/15 < 50%).

Table 4.1: Mean Absolute Error Across 15 Attribute Shelf and CBC Holdouts Including the None

		Shelf-Looking Holdouts				CBC-Looking Holdouts						
		FP	PP	MAE		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	MAE
Including the None	CBC FP	2.4%	3.1%	2.7%	CBC FP	3.1%	1.7%	5.3%	1.9%	3.0%	6.3%	3.6%
	CBC PP	2.0%	2.6%	2.3%	CBC PP	6.6%	10.1%	8.2%	8.1%	5.8%	4.5%	7.2%
	ACBC FP	1.5%	2.8%	2.2%	ACBC FP	2.6%	4.3%	4.1%	3.2%	3.2%	3.0%	3.4%
	ACBC PP	1.3%	1.7%	1.5%	ACBC PP	3.2%	1.8%	1.5%	3.3%	2.0%	5.5%	2.9%

Table 4.2: Mean Absolute Error Across 15 Attribute Shelf and CBC Holdouts Excluding the None

		Shelf-Looking Holdouts				CBC-Looking Holdouts						
		FP	PP	MAE		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	MAE
Dropping the None	CBC FP	3.0%	3.9%	3.4%	CBC FP	3.4%	4.5%	8.1%	2.6%	4.2%	8.3%	5.2%
	CBC PP	2.0%	1.9%	2.0%	CBC PP	11.3%	12.3%	15.9%	11.2%	9.9%	6.5%	11.2%
	ACBC FP	2.1%	2.4%	2.3%	ACBC FP	1.7%	12.0%	8.7%	7.2%	6.6%	8.1%	7.4%
	ACBC PP	2.1%	2.5%	2.3%	ACBC PP	4.1%	5.0%	11.4%	8.4%	1.6%	10.7%	6.9%

The None Parameter

Again, we see that those in the ACBC sample cells are significantly more likely to choose the None in the holdout choices (i.e., respondents would not buy any product on the shelf) (Table 4.3). In addition, both PP methods, PP ACBC and PP CBC, seem to align more closely to shelf behavior as seen in the survey tool.

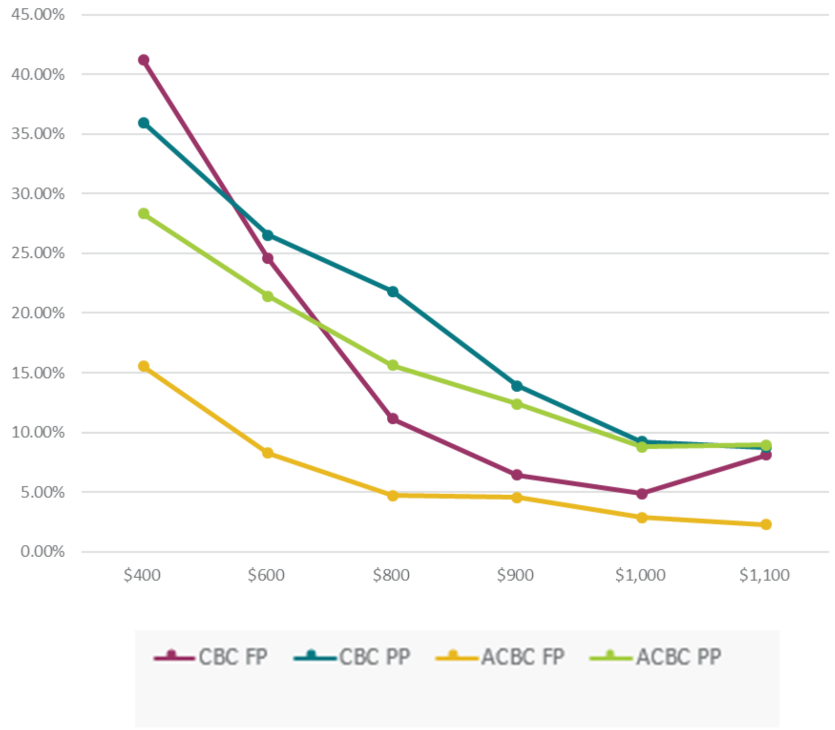
Table 4.3: None Proportions within the 15 Attribute Experiments

	CBC FP	CBC PP	ACBC FP	ACBC PP
PP Shelf Holdout				
None simulated (Exp=1)	44%	22%	62%	38%
None actual	21%	13%	25%	31%
Difference	+23%	+9%	+37%	+6%
FP Shelf Holdout				
None simulated (Exp=1)	37%	10%	48%	28%
None actual	18%	14%	28%	29%
Difference	+19%	-5%	+20%	-2%

Price Sensitivity

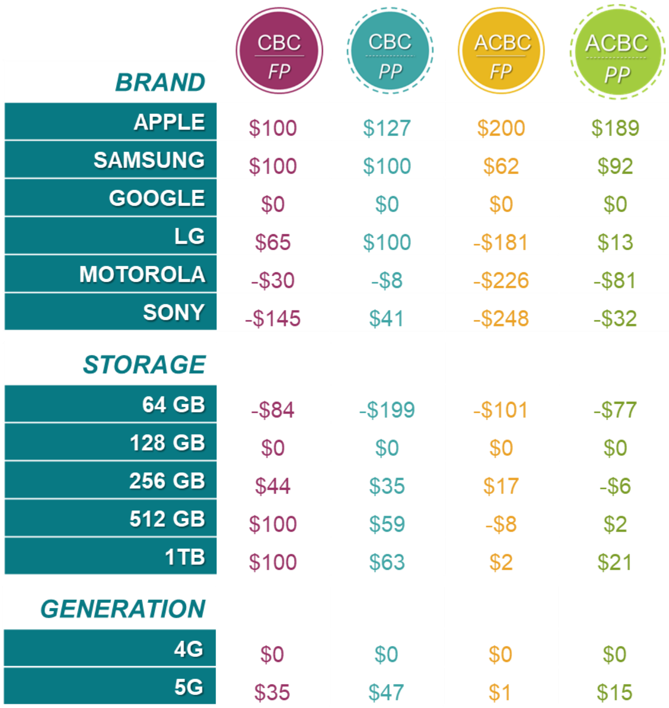
Figure 4.4 shows the share of preference estimated from each model when only changing price. FP ACBC is the most conservative when predicting choice.

Figure 4.4: Price Sensitivity of 15 Attributes by Methodology



Willingness to Pay (WTP)

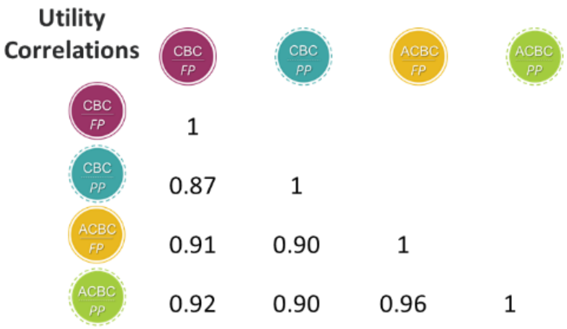
Figure 4.5: WTP for Brand, Storage, and Generation Attributes by Methodology



BRAND	CBC FP	CBC PP	ACBC FP	ACBC PP
APPLE	$100	$127	$200	$189
SAMSUNG	$100	$100	$62	$92
GOOGLE	$0	$0	$0	$0
LG	$65	$100	-$181	$13
MOTOROLA	-$30	-$8	-$226	-$81
SONY	-$145	$41	-$248	-$32
STORAGE				
64 GB	-$84	-$199	-$101	-$77
128 GB	$0	$0	$0	$0
256 GB	$44	$35	$17	-$6
512 GB	$100	$59	-$8	$2
1TB	$100	$63	$2	$21
GENERATION				
4G	$0	$0	$0	$0
5G	$35	$47	$1	$15

It is difficult to tell from these estimates which methodology is most accurate. This could be due to aggregating WTP values or differences in methodology. However, we can see that when comparing utility correlations, the ACBC methods are more closely aligned (Figure 4.6).

Figure 4.6: Correlations of Aggregate Utilities by Methodology

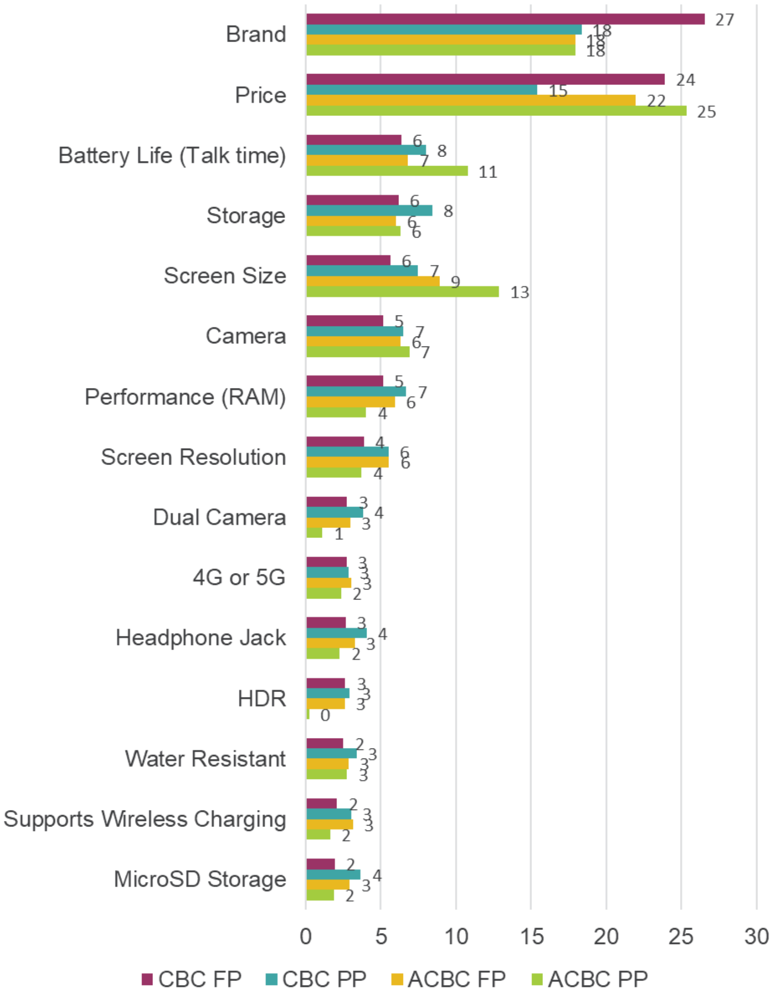


Utility Correlations	CBC FP	CBC PP	ACBC FP	ACBC PP
CBC FP	1			
CBC PP	0.87	1		
ACBC FP	0.91	0.90	1	
ACBC PP	0.92	0.90	0.96	1

Importance Scores

With PP ACBC, a dropped attribute's importance is set to 0. Therefore, it is not surprising that PP ACBC has very steep importance scores, similar to FP CBC. Similar to the 10 attribute results, the ACBC data calculates price more important than brand, while the CBC data calculates brand as more important than price (Figure 4.7).

Figure 4.7: Importance Scores by Methodology



Respondent Preference

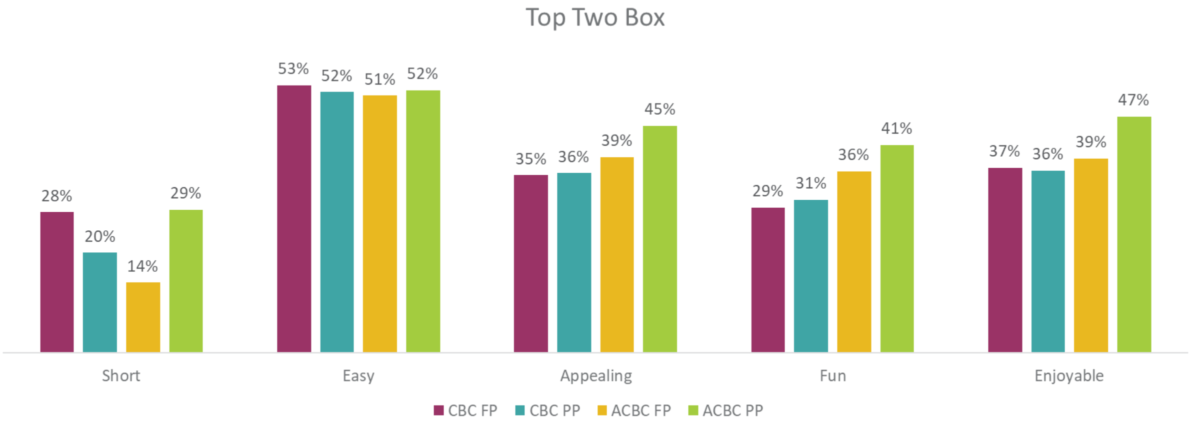
FP CBC is the quickest exercise to complete, while FP ACBC is the longest (Figure 4.8). All methods have an equal drop-off rate. PP CBC has the highest proportion of respondents admitting to cheating and the highest proportion of bad data. Again, this could be due to the low percentage of attributes shown per screen, and a harder cognitive burden (7 attributes/ 15 attributes).

Figure 4.8: Respondent Statistics for 15 Attribute Experiments

CBC FP	CBC PP	ACBC FP	ACBC PP
MEDIAN TIME – 5.9 MIN	MEDIAN TIME – 7.2 MIN	MEDIAN TIME – 14.9 MIN	MEDIAN TIME – 10.4 MIN
DROP OFF – 6.4%	DROP OFF – 6.2%	DROP OFF – 6.3%	DROP OFF – 6.2%
ADMIT CHEATING – 5.0%	ADMIT CHEATING – 7.9%	ADMIT CHEATING – 5.9%	ADMIT CHEATING – 4.8%
BAD DATA* – 6.5%	BAD DATA* – 11.5%	BAD DATA* – 9.5%	BAD DATA* – 8.8%

When rating the different methodologies, PP ACBC appears to be the most preferred (Figure 4.9).

Figure 4.9: Respondent Top Two Agreement on 15 Attribute Survey Experience



15 Attribute Conclusion

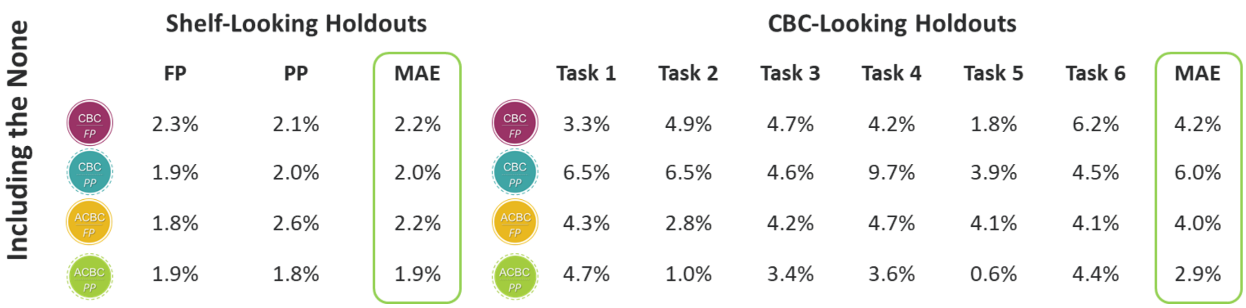
PP ACBC seems to edge out FP CBC, particularly from the respondent perspective. From a modeling perspective to predict holdout choice shares, FP CBC and PP ACBC are likely equal, however more research should be done into the significant differences in the None parameter estimates and the impact that has on predicting actual market share.

20 Attribute Story

Model Validity

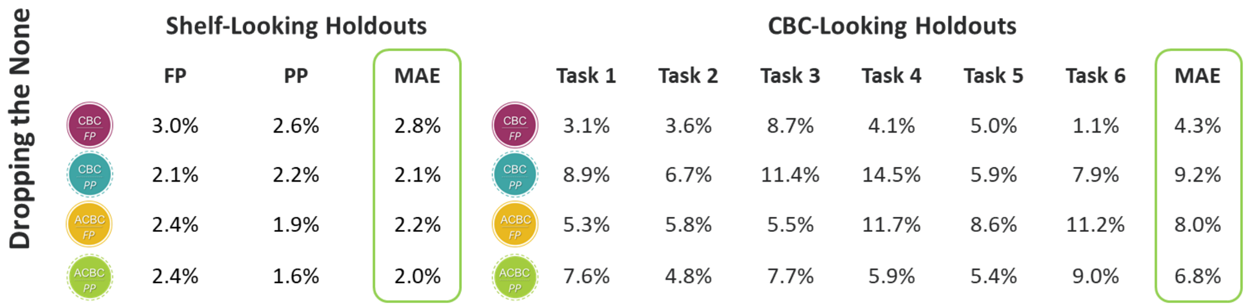
Similar to the 15 attribute findings, Table 5.1 shows that when including the None option in simulations, ACBC has the lowest MAE, particularly PP ACBC. After dropping the None, the MAE for the Adaptive methods are more in line with FP CBC (Table 5.2). Again, PP CBC has the highest MAE, perhaps due to the number of attributes shown out of the total attributes modeled (7/20 ~35%). (Again, we should note that the CBC-looking holdout tasks would be expected to bias this measure of predictive validity in favor of the CBC methods.)

Table 5.1: Mean Absolute Error Across 20 Attribute Shelf and CBC Holdouts Including the None



Including the None	Shelf-Looking Holdouts				CBC-Looking Holdouts						
	FP	PP	MAE		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	MAE
CBC FP	2.3%	2.1%	2.2%	CBC FP	3.3%	4.9%	4.7%	4.2%	1.8%	6.2%	4.2%
CBC PP	1.9%	2.0%	2.0%	CBC PP	6.5%	6.5%	4.6%	9.7%	3.9%	4.5%	6.0%
ACBC FP	1.8%	2.6%	2.2%	ACBC FP	4.3%	2.8%	4.2%	4.7%	4.1%	4.1%	4.0%
ACBC PP	1.9%	1.8%	1.9%	ACBC PP	4.7%	1.0%	3.4%	3.6%	0.6%	4.4%	2.9%

Table 5.2: Mean Absolute Error Across 20 Attribute Shelf and CBC Holdouts Excluding the None



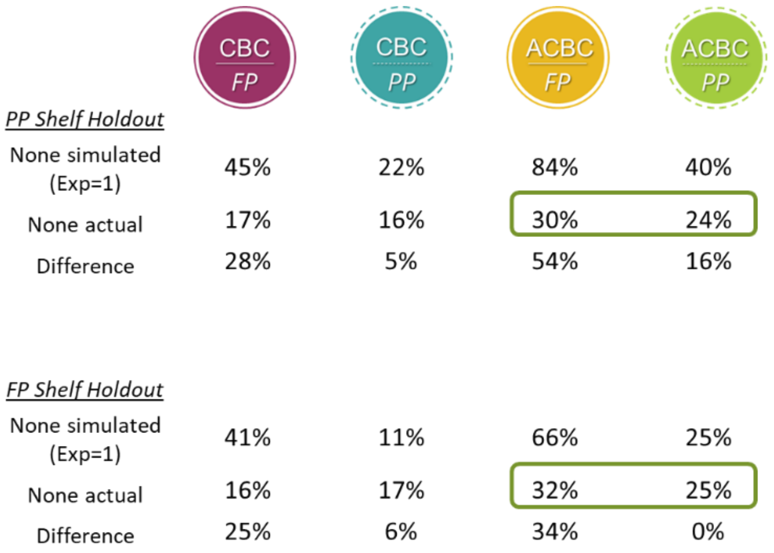
Dropping the None

	Shelf-Looking Holdouts				CBC-Looking Holdouts						
	FP	PP	MAE		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	MAE
CBC FP	3.0%	2.6%	2.8%	CBC FP	3.1%	3.6%	8.7%	4.1%	5.0%	1.1%	4.3%
CBC PP	2.1%	2.2%	2.1%	CBC PP	8.9%	6.7%	11.4%	14.5%	5.9%	7.9%	9.2%
ACBC FP	2.4%	1.9%	2.2%	ACBC FP	5.3%	5.8%	5.5%	11.7%	8.6%	11.2%	8.0%
ACBC PP	2.4%	1.6%	2.0%	ACBC PP	7.6%	4.8%	7.7%	5.9%	5.4%	9.0%	6.8%

The None Parameter

Again, we see that those in the ACBC sample cells are significantly more likely to choose the None (i.e., They would not buy any product on the shelf) (Table 5.3). Also, both PP methods, PP ACBC and PP CBC, seem to align more closely to shelf behavior demonstrated in the survey tool.

Table 5.3: None Proportions within the 20 Attribute Experiments

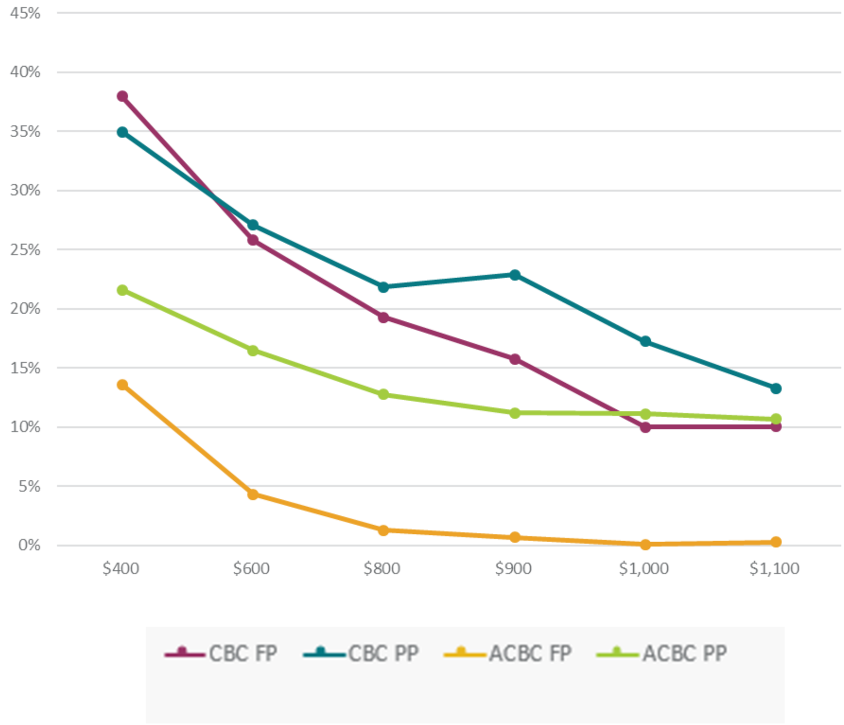


	CBC FP	CBC PP	ACBC FP	ACBC PP
PP Shelf Holdout				
None simulated (Exp=1)	45%	22%	84%	40%
None actual	17%	16%	30%	24%
Difference	28%	5%	54%	16%
FP Shelf Holdout				
None simulated (Exp=1)	41%	11%	66%	25%
None actual	16%	17%	32%	25%
Difference	25%	6%	34%	0%

Price Sensitivity

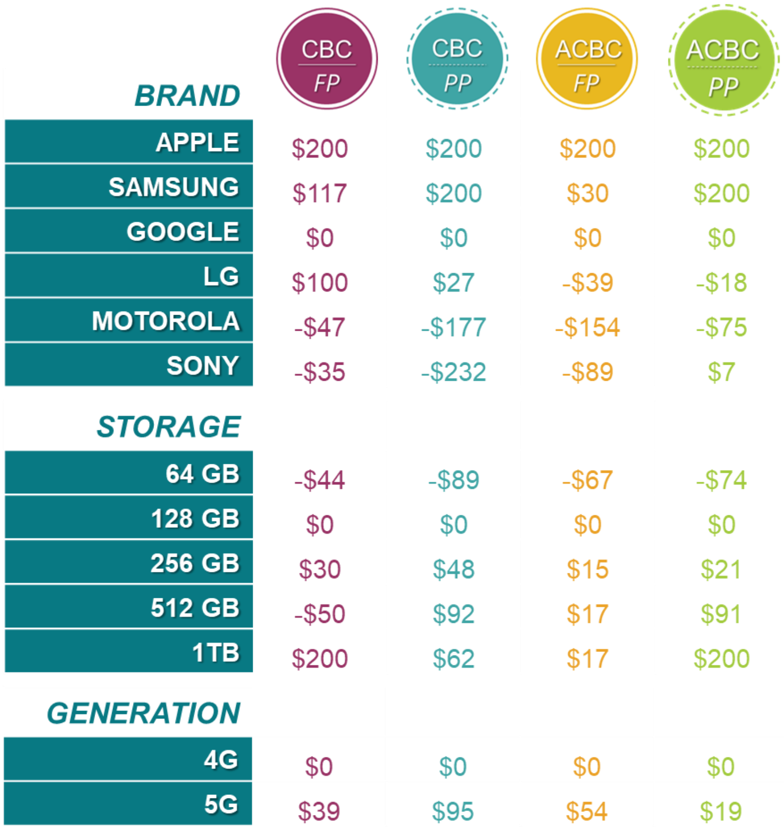
Figure 5.4 shows the share of preference estimated from each model when only changing price. Again, FP ACBC is the most conservative when predicting choice.

Figure 5.4: Price Sensitivity of 20 Attributes by Methodology



Willingness to Pay (WTP)

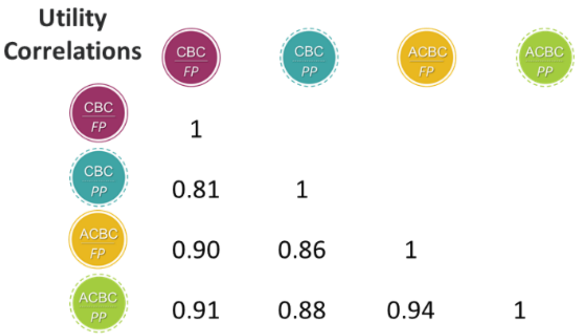
Figure 5.5—WTP for Brand, Storage, and Generation Attributes by Methodology



BRAND	CBC FP	CBC PP	ACBC FP	ACBC PP
APPLE	$200	$200	$200	$200
SAMSUNG	$117	$200	$30	$200
GOOGLE	$0	$0	$0	$0
LG	$100	$27	-$39	-$18
MOTOROLA	-$47	-$177	-$154	-$75
SONY	-$35	-$232	-$89	$7
STORAGE				
64 GB	-$44	-$89	-$67	-$74
128 GB	$0	$0	$0	$0
256 GB	$30	$48	$15	$21
512 GB	-$50	$92	$17	$91
1TB	$200	$62	$17	$200
GENERATION				
4G	$0	$0	$0	$0
5G	$39	$95	$54	$19

When testing 20 attributes, the ACBC methods more closely align on WTP, even when considering PP ACBC did not force price into the exercise. We can also see this pattern when comparing utility correlations (Figure 5.6).

Figure 5.6: Correlations of Aggregate Utilities by Methodology



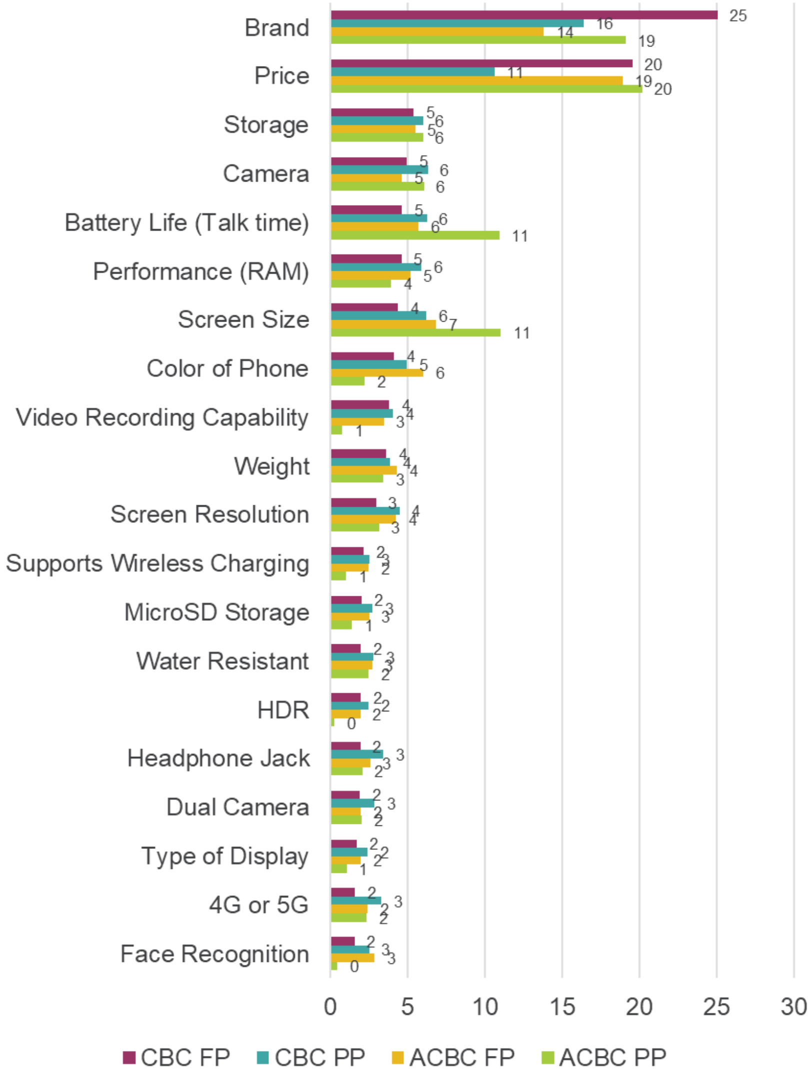
Utility Correlations	CBC FP	CBC PP	ACBC FP	ACBC PP
CBC FP	1			
CBC PP	0.81	1		
ACBC FP	0.90	0.86	1	
ACBC PP	0.91	0.88	0.94	1

Importance Scores

Similar to the 10 and 15 attribute results, the ACBC data calculates price more important than brand, while the CBC data calculates brand as more important than price (Figure 5.7).

FP CBC has the steepest importance scores, as expected, while PP CBC has the flattest, also expected.

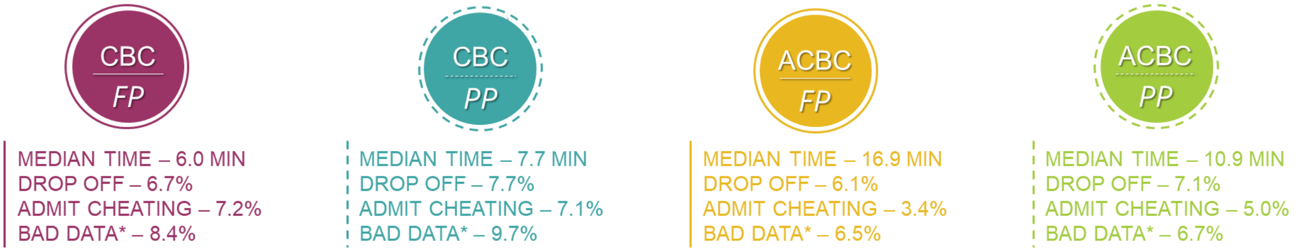
Figure 5.7: Importance Scores by Methodology



Respondent Preference

FP CBC is the quickest exercise to complete, while FP ACBC is the longest (Figure 5.8). The drop-off rates for the PP methods are slightly higher than FP. The CBC methods have the highest proportion of cheaters and bad data.

Figure 5.8: Respondent Statistics for 20 Attribute Experiments



CBC FP
MEDIAN TIME – 6.0 MIN
DROP OFF – 6.7%
ADMIT CHEATING – 7.2%
BAD DATA* – 8.4%

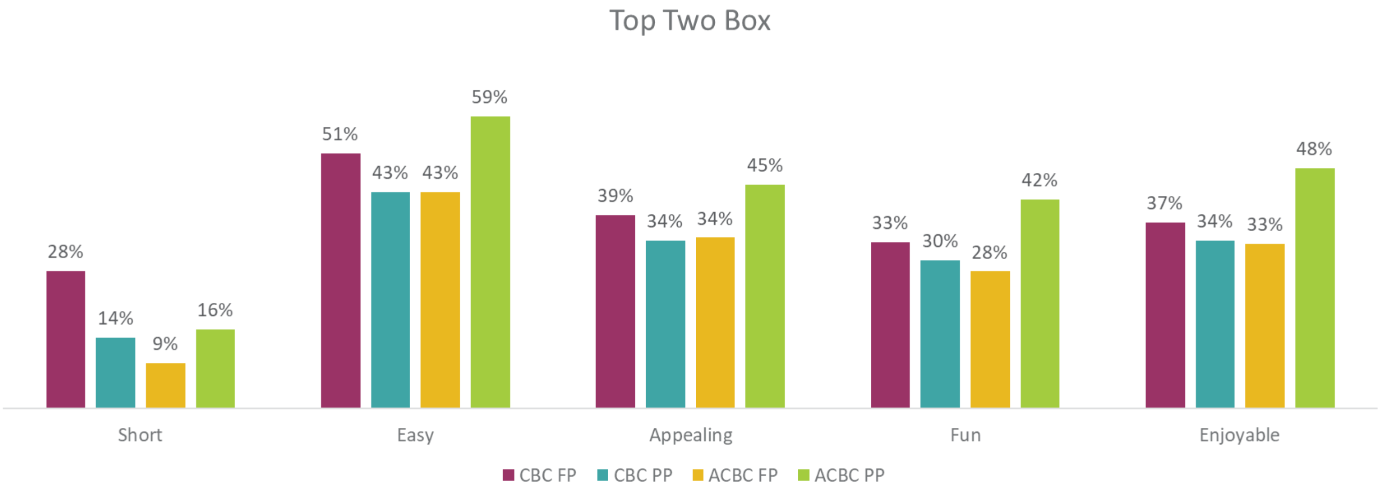
CBC PP
MEDIAN TIME – 7.7 MIN
DROP OFF – 7.7%
ADMIT CHEATING – 7.1%
BAD DATA* – 9.7%

ACBC FP
MEDIAN TIME – 16.9 MIN
DROP OFF – 6.1%
ADMIT CHEATING – 3.4%
BAD DATA* – 6.5%

ACBC PP
MEDIAN TIME – 10.9 MIN
DROP OFF – 7.1%
ADMIT CHEATING – 5.0%
BAD DATA* – 6.7%

When rating the different methodologies, PP ACBC is the clear winner after which respondents tend to prefer FP CBC (Figure 5.9).

Figure 5.9: Respondent Top Two Agreement on 20 Attribute Survey Experience



20 Attribute Conclusion

PP ACBC increases its lead on FP CBC, in comparison to the 15-attribute story, both from a model and respondent perspective.

Overall Recommendations

Even though we noticed differences between the methods across the different attributes tested, all these methods have been thoroughly researched and tested to handle larger numbers of attributes. In this specific design, when testing up to 10 attributes, the researcher seems to have more flexibility in the methodology chosen as the models and respondent experience is comparable. However, as we increase to 15+ attributes, we did recognize a clear winner in the form of PP ACBC. Although PP ACBC performs similarly to the FP CBC with regard to model accuracy, we shouldn't overlook the fact that model accuracy was gauged in terms of ability to predict shares for CBC-looking holdouts, which naturally should have favored the FP CBC method. Moreover, respondent satisfaction scores give PP ACBC significant credibility to use as attributes increase.

Considerations for Future Research

A lot of decisions were made during the course of the research that lead to the results of this paper. One of those decisions was to not display “brand” and “price” on each of the partial profile methods. In the future, others may consider forcing these attributes into all product profiles, if appropriate for the product category. The authors are extremely interested in understanding the impact of this decision and whether or not the model accuracy is significantly improved when accounting for this.

Another consideration is to further explore the high None parameter estimated in the ACBC models and whether there is a psychological effect (building your own product first makes you less likely to like other products) or a design effect (computing the None parameter from screening tasks is not the same as what respondents would do in the real world).

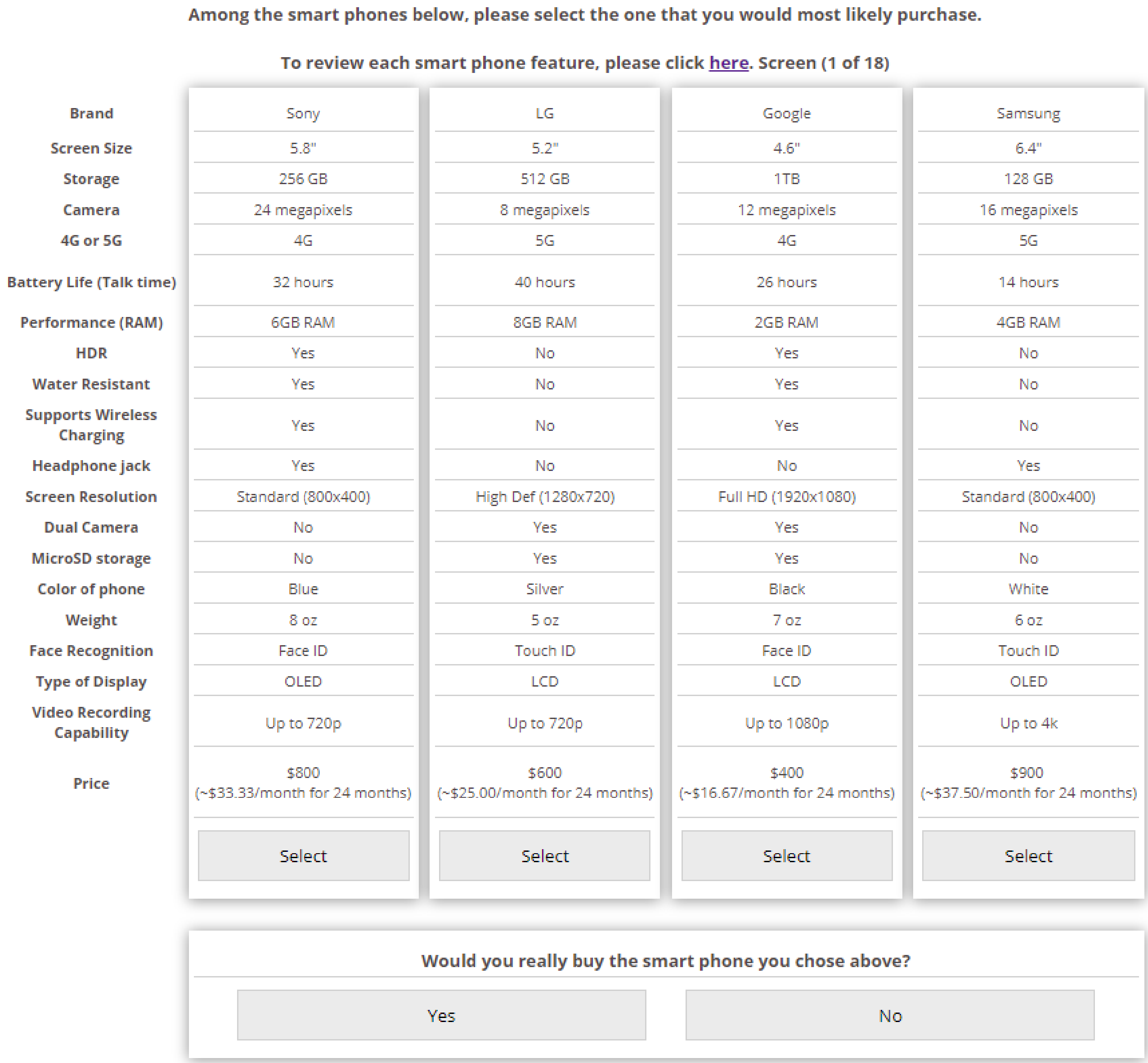
Megan Peitz

Mike Serpetti

Dan Yardley

APPENDIX

Full Profile CBC Example (20 Attribute)



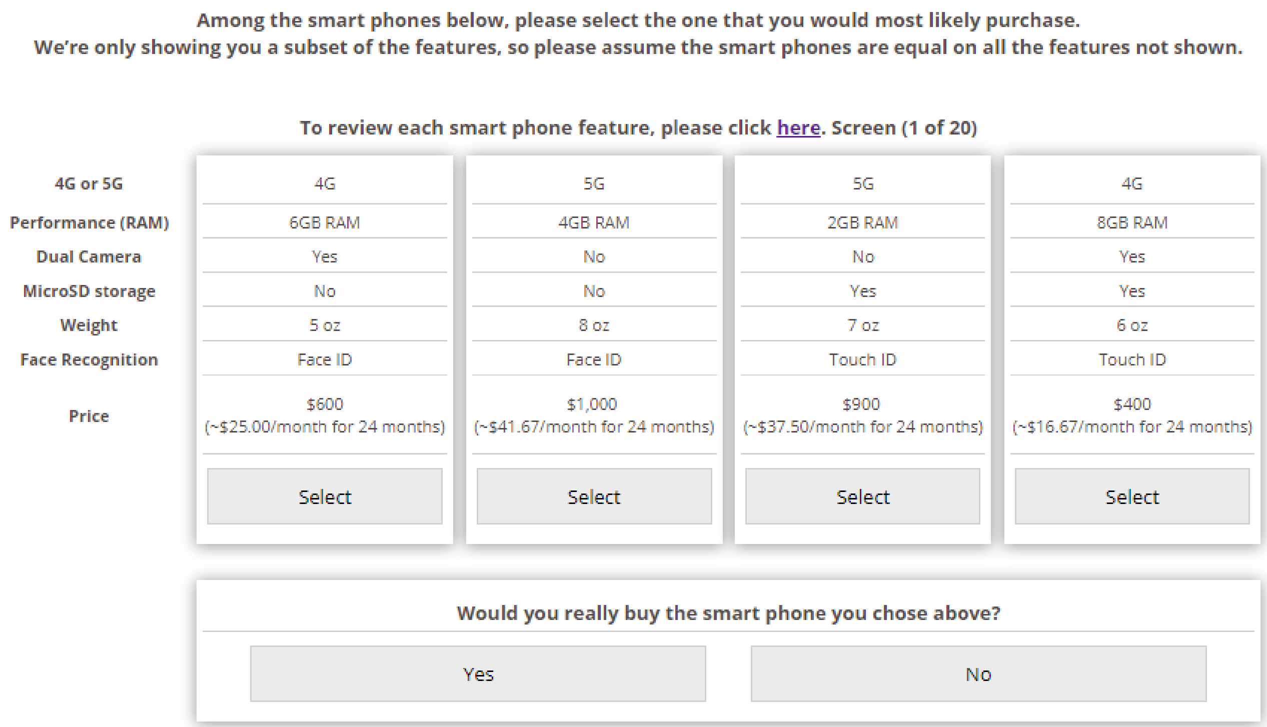
Among the smart phones below, please select the one that you would most likely purchase.

To review each smart phone feature, please click here. Screen (1 of 18)

Brand	Sony	LG	Google	Samsung
Screen Size	5.8"	5.2"	4.6"	6.4"
Storage	256 GB	512 GB	1TB	128 GB
Camera	24 megapixels	8 megapixels	12 megapixels	16 megapixels
4G or 5G	4G	5G	4G	5G
Battery Life (Talk time)	32 hours	40 hours	26 hours	14 hours
Performance (RAM)	6GB RAM	8GB RAM	2GB RAM	4GB RAM
HDR	Yes	No	Yes	No
Water Resistant	Yes	No	Yes	No
Supports Wireless Charging	Yes	No	Yes	No
Headphone jack	Yes	No	No	Yes
Screen Resolution	Standard (800x400)	High Def (1280x720)	Full HD (1920x1080)	Standard (800x400)
Dual Camera	No	Yes	Yes	No
MicroSD storage	No	Yes	Yes	No
Color of phone	Blue	Silver	Black	White
Weight	8 oz	5 oz	7 oz	6 oz
Face Recognition	Face ID	Touch ID	Face ID	Touch ID
Type of Display	OLED	LCD	LCD	OLED
Video Recording Capability	Up to 720p	Up to 720p	Up to 1080p	Up to 4k
Price	$800 (~$33.33/month for 24 months)	$600 (~$25.00/month for 24 months)	$400 (~$16.67/month for 24 months)	$900 (~$37.50/month for 24 months)
	Select	Select	Select	Select

Would you really buy the smart phone you chose above?

Partial Profile CBC Example (20 Attribute)



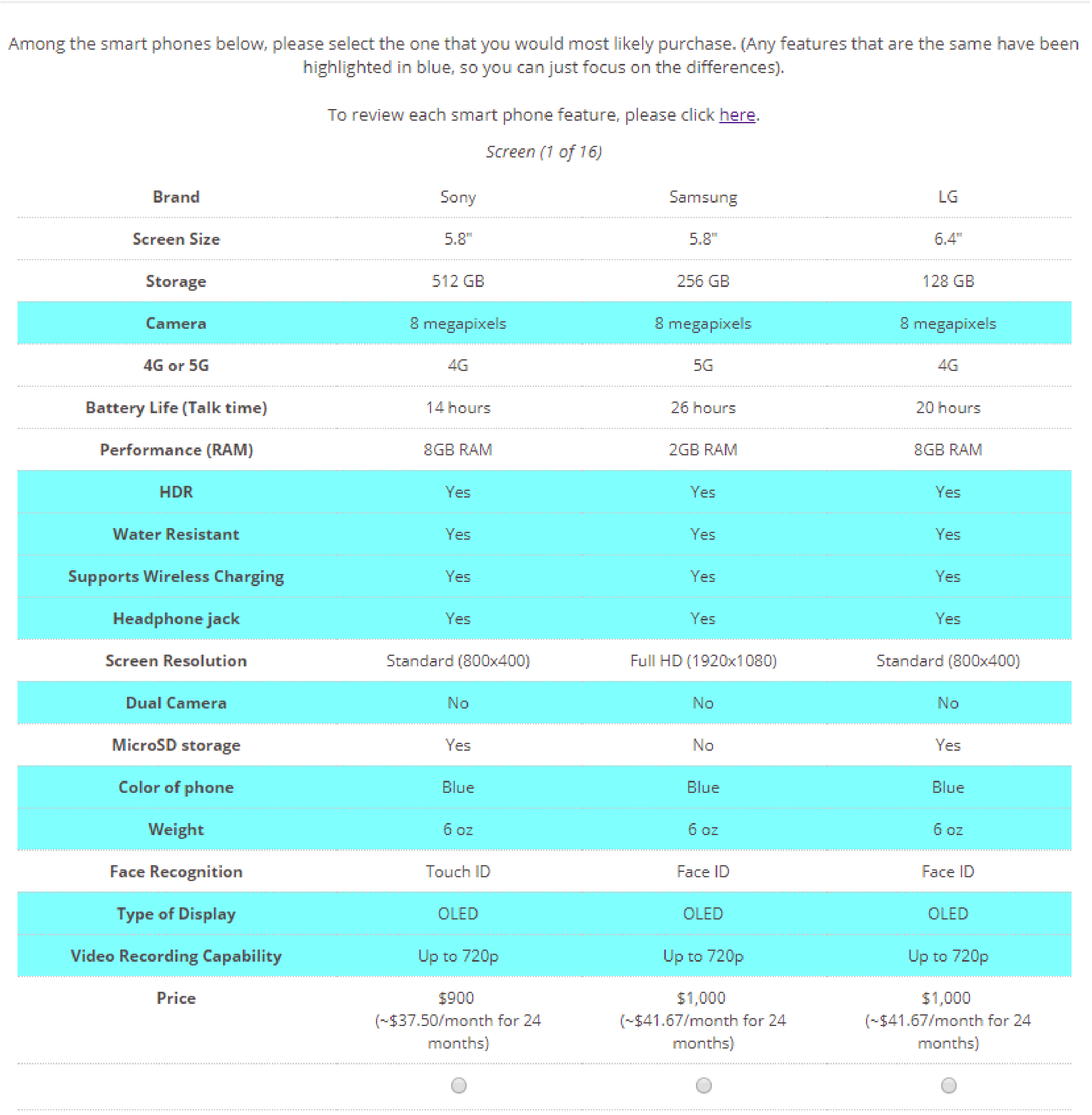
**Among the smart phones below, please select the one that you would most likely purchase.
We're only showing you a subset of the features, so please assume the smart phones are equal on all the features not shown.**

To review each smart phone feature, please click here. Screen (1 of 20)

4G or 5G	4G	5G	5G	4G
Performance (RAM)	6GB RAM	4GB RAM	2GB RAM	8GB RAM
Dual Camera	Yes	No	No	Yes
MicroSD storage	No	No	Yes	Yes
Weight	5 oz	8 oz	7 oz	6 oz
Face Recognition	Face ID	Face ID	Touch ID	Touch ID
Price	$600 (~$25.00/month for 24 months)	$1,000 (~$41.67/month for 24 months)	$900 (~$37.50/month for 24 months)	$400 (~$16.67/month for 24 months)

Would you really buy the smart phone you chose above?

Full Profile ACBC Example (20 Attribute)



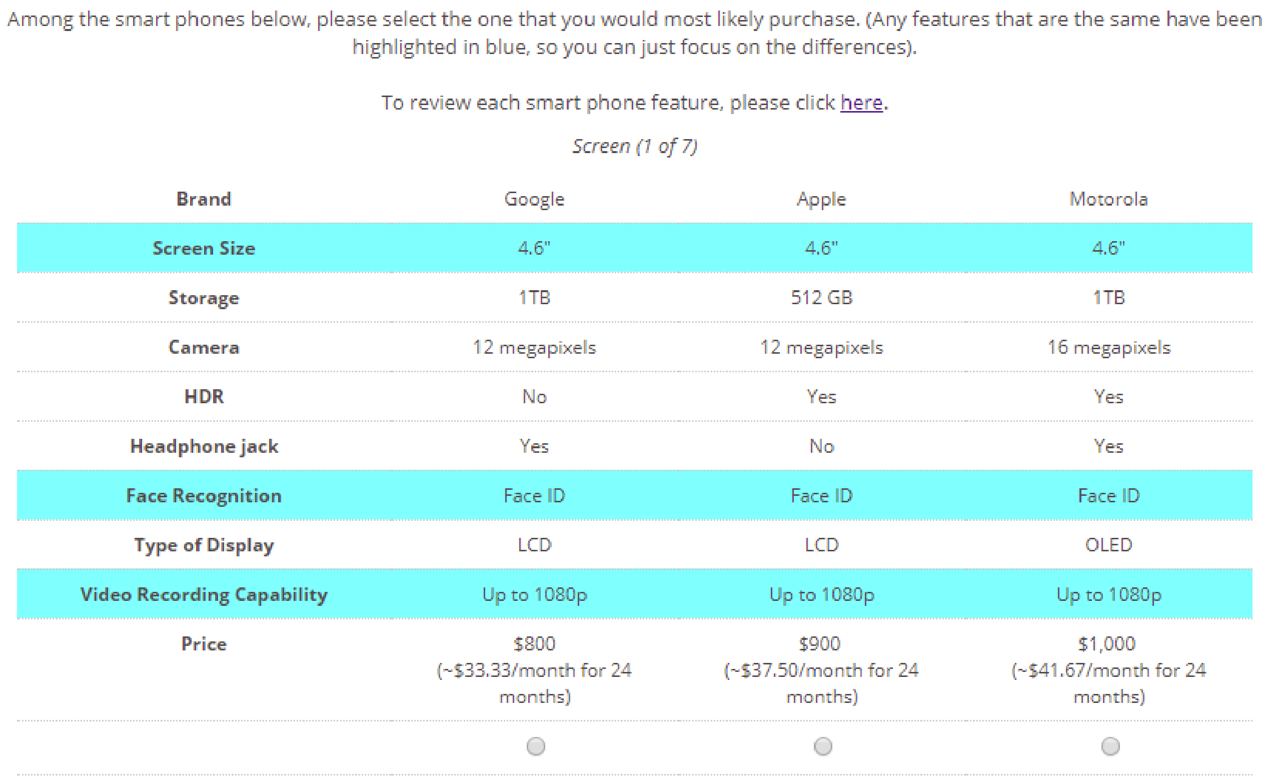
Among the smart phones below, please select the one that you would most likely purchase. (Any features that are the same have been highlighted in blue, so you can just focus on the differences).

To review each smart phone feature, please click here.

Screen (1 of 16)

Brand	Sony	Samsung	LG
Screen Size	5.8"	5.8"	6.4"
Storage	512 GB	256 GB	128 GB
Camera	8 megapixels	8 megapixels	8 megapixels
4G or 5G	4G	5G	4G
Battery Life (Talk time)	14 hours	26 hours	20 hours
Performance (RAM)	8GB RAM	2GB RAM	8GB RAM
HDR	Yes	Yes	Yes
Water Resistant	Yes	Yes	Yes
Supports Wireless Charging	Yes	Yes	Yes
Headphone jack	Yes	Yes	Yes
Screen Resolution	Standard (800x400)	Full HD (1920x1080)	Standard (800x400)
Dual Camera	No	No	No
MicroSD storage	Yes	No	Yes
Color of phone	Blue	Blue	Blue
Weight	6 oz	6 oz	6 oz
Face Recognition	Touch ID	Face ID	Face ID
Type of Display	OLED	OLED	OLED
Video Recording Capability	Up to 720p	Up to 720p	Up to 720p
Price	$900 (~$37.50/month for 24 months)	$1,000 (~$41.67/month for 24 months)	$1,000 (~$41.67/month for 24 months)
	○	○	○

Partial Profile ACBC Example (20 Attribute)



Among the smart phones below, please select the one that you would most likely purchase. (Any features that are the same have been highlighted in blue, so you can just focus on the differences).

To review each smart phone feature, please click here.

Screen (1 of 7)

Brand	Google	Apple	Motorola
Screen Size	4.6"	4.6"	4.6"
Storage	1TB	512 GB	1TB
Camera	12 megapixels	12 megapixels	16 megapixels
HDR	No	Yes	Yes
Headphone jack	Yes	No	Yes
Face Recognition	Face ID	Face ID	Face ID
Type of Display	LCD	LCD	OLED
Video Recording Capability	Up to 1080p	Up to 1080p	Up to 1080p
Price	$800 (~$33.33/month for 24 months)	$900 (~$37.50/month for 24 months)	$1,000 (~$41.67/month for 24 months)
	○	○	○

References

Elrod, T. and Chrzan, K. (1994), Partial Profile Conjoint Analysis: a choice-based approach for handling large numbers of attributes. Faculty of Business, University of Alberta, Canada.

Elrod, T. and Chrzan, K. (1994), Choice-based Approach for Large Numbers of Attributes, Marketing News, 29, 20–30.

Green, Paul E and Srinivasan, V (1978). Conjoint Analysis in Consumer Research: Issues and Outlook, *Journal of Consumer Research*, 5(2), 103–123.

Kurz, Peter. A Comparison between Adaptive Choice Based Conjoint, Partial Profile Choice Based Conjoint and Choice Based Conjoint. 2009. SKIM Conference, Prague.

Orme, B. and Johnson, R. (2007) A New Approach to Adaptive CBC. https://sawtoothsoftware.com/download/techpap/acbc10.pdf

WHAT SOUTH AFRICAN MEDICAL STUDENTS' VALUE IN A RURAL INTERNSHIP: A DISCRETE CHOICE EXPERIMENT

MARIA JOSE
UNIVERSITY OF CAPE TOWN

INTRODUCTION

The World Health Organization has identified the Health workforce as a critical building block of a functional health system (World Health Organization, 2010). However, inequality in healthcare access between urban and rural areas has historically resulted in poor health outcomes in developing countries, especially in Africa. Rural medical practice is unpopular due to social and cultural isolation and lack of infrastructure (Lagarde and Blaauw, 2014). In South Africa, a constitutional democracy since 1994, recruitment policies have been enacted to increase the health workforce working in government health facilities in rural areas.

BACKGROUND

In South Africa a medical graduate's journey from high school to an independent medical practitioner is as follows: once completing high school in Grade 12, the student would be enrolled in a public university (there are no private medical schools in South Africa) for a period of 6 years of undergraduate study and would graduate with a medical degree. Thereafter they would be required to serve 2 years of paid internship at a government health facility in the country under supervision before they work independently for a 1-year community service period, also at a government facility. It is only once they have completed this 9-year process that the student is registered by the Health Profession's Council to practice medicine in the country as a General Practitioner. It is the career transition from graduate to intern which is the focus of this study as at this juncture medical graduates can choose which government facility to work at during their internship period, and historically rural facilities are an unpopular choice. Rural as used in this study is defined as "an area more than two hours' travel by road from the nearest urban center," whereas urban is defined as "a center with a population of more than 250,000 people" (Burch et al., 2017).

RESEARCH QUESTIONS

- What are the rural facility attributes that final-year medical students value?
- When faced with conflicting rural facility attributes, what tradeoffs are final-year medical students willing to make?
- In monetary terms how much do final-year medical students value each facility attribute?
- Are rural facility attribute preferences influenced by gender, career intention, and undergraduate rural medicine exposure?

Ethics approval for this study was granted by the University of Cape Town Human Research Ethics Committee and Department of Student Affairs.

METHODOLOGY

The study included a Literature review, focus group discussions (FGD), and a discrete choice experiment (DCE or CBC).

The Literature Review consisted of 25 articles pertaining to current recruitment policies in South Africa and healthcare worker job attribute preferences in other countries. Currently South Africa has instituted a mandatory, paid 1-year community service for all medical graduates, however this has proven to be a costly initiative and contributes to high staff turnover that further destabilises fragile rural health facilities (Dambisya, 2007). Another program known as the Mandela-Castro program invests in training rural-origin students in Cuba under the condition that they work in rural areas upon graduation. This too is costly and the clinical training of such graduates is not locally relevant due to the different burden of disease in Cuba versus South Africa (Dambisya, 2007). Finally, both rural allowance and on-site housing provision have been costly and have not proven to be effective in recruitment of medical graduates to rural areas. Interestingly, there is evidence to suggest that identifying and training rural-origin students can have lasting benefits for rural healthcare service provision depending on the context.

In total 3 FGDs were conducted with 15 participants before saturation was reached. During the focus groups the following attribute was identified by literature reviews but then discarded as they are personal, not facility factors: proximity of the healthcare facility to the medical graduate's partner or child's school/work. During the focus groups the following attributes were identified by literature reviews but then discarded as they were not supported in the FGDs: access to Wi-Fi, in-service training, and socialization among colleagues. Attributes that were mentioned in literature and supported in FGDs were included in the DCE (Table 1): rural allowance, housing provision, physical safety, and availability of basic resources at facility. Attributes which were not mentioned in other literature but came out strongly in the FGDs: the provision of personal protective equipment against occupational tuberculosis exposure, the extent of practical experience ingrained at the facility, and the seniority of the supervisor, with Medical Officer being the most junior and Consultant being the most senior.

Attributes and levels were used to form hypothetical job postings in a Traditional CBC design in Sawtooth Software Lighthouse studio 9.6.1. Design settings: Complete Enumeration, 2 Concepts per Task which were discrete choice response types. Unlabeled with no opt-out option or fixed tasks. 15 Random Tasks and 12 demographic questions were included, and 1 questionnaire version, 1 design seed with no attribute randomization and no concept sorting.

No opt out option was chosen to replicate real-life decision-making as graduates have limited placement choices; if a graduate declined a placement, they would not be able to practice medicine. The questionnaire was piloted with 25 students and the wording improved for clarity.

Table 1: List of attributes and levels used in the DCE

Attribute	Attribute Level and Description
Supervision	-Supervised by Medical Officer -Supervised by Registrar -Supervised by Consultant
Rural Allowance	-R4,000 per month -R4,340 per month (8% increase) -R4,800 per month (20% increase)
Accommodation	-Rent private accommodation -Provided with subsidised doctors quarters on hospital premises
Resources	-Daily stock out of gloves, syringes and suture packs -Gloves, syringes and suture packs available daily
Practical Experience	-Limited to filling out forms and taking bloods -Includes filling out forms, take bloods and doing procedures e.g., lumbar punctures
Hospital Safety	-There have been few reports of theft, hijacking and protests in and around the hospital in the past year -There is a high level of crime in and around the hospital with many reports of theft, hijacking and protests in the past year
Occupational Hazard	-No tuberculosis masks available in the hospital -Poorly fitting tuberculosis masks always available -Correctly sized tuberculosis masks always available

All final-year medical students at universities across South Africa who will be applying for internship placement were included and through purposive sampling were invited to participate through email. Data was captured and processed in Sawtooth Software Lighthouse studio V9.6.1 accessed via an academic grant.

RESULTS

The final sample size was 357 with 221 completed and 136 incomplete questionnaires (61.9% completion rate). Table 2 shows the demographics which, unsurprisingly, represent the general medical student body; majority being females in their mid-twenties hailing from urban areas, unmarried without child dependents, with some rural medicinal exposure, intending to specialise.

Table 2: Demographic Factors

Variables		Number (%)
Age	Mean 23.7 years	—
Gender	Male	70 (31.7)
	Female	145 (65.6)
	Non-Conforming	6 (2.7)
Province of origin	Western Cape	75 (33.9)
	Gauteng	62 (28.1)
	Free State	0 (0.0)
	North West	3 (1.4)
	Eastern Cape	21 (9.5)
	Kwa-Zulu Natal	42 (19.0)
	Mpumalanga	8 (3.6)
	Limpopo	7 (3.2)
	Northern Cape	3 (1.4)
Area of origin	Rural (village/farm)	15 (6.8)
	Informal settlement	6 (2.7)
	Urban (formal structure in suburb/township)	200 (90.5)
Marital status	Single, never married	205 (92.8)
	Married	16 (7.2)
	Widowed	0 (0.0)
	Divorced/separated	0 (0.0)
Child dependents	Yes	3 (1.4)
	No	218 (98.6)
Undergrad exposure rural Med	Yes	127 (57.5)
	No	94 (42.5)
Provincial bursary	Yes	46 (20.8)
	No	175 (79.2)
Cuban trained student	Yes	7 (3.2)
	No	214 (96.8)
Intention to intern	Yes	217 (98.2)
	No	4 (1.8)
Career intention	General Practice	12 (5.5)
	Specialisation	122 (56.0)
	I don’t know/ undecided	78 (35.8)
	Other	5 (2.3)
	Did not intend to complete internship	4 (1.8)

Table 3 shows the average of the hierarchical Bayes utilities of the attribute levels which indicate that advanced practical experience, correctly sized tuberculosis masks, and hospital safety were the utility maximizing preferences. Interestingly neither rural allowance increases, nor housing provision provided as high average utilities.

Table 3: Hierarchical Bayes Estimation of Average Utility

Attributes	**Average Utilities (Zero-Centered)**	**Standard Deviation**
Supervised by Medical Officer	-12.6	35.6
Supervised by Registrar	-11.0	21.3
Supervised by Consultant	23.6	31.2
Practical experience is limited to filling out forms and taking bloods	-76.3	55.8
Practical experience includes filling out forms, take bloods and doing procedures e.g., lumbar punctures	76.3	55.8
Daily stock out of gloves, syringes and suture packs	-52.8	34.1
Gloves, syringes and suture packs available daily	52.8	34.1
Rural allowance R4000 (current level)	-36.9	22.8
Rural allowance R4340 (8% increase)	14.9	18.2
Rural allowance R4800 (20% increase)	22.0	28.3
There have been few reports of theft, hijacking and protests in and around the hospital in the past year	61.0	54.6
There is a high level of crime in and around the hospital with many reports of theft, hijacking and protests in the past year.	-61.0	54.6
No tuberculosis masks available in the hospital	-63.2	33.3
Poorly fitting tuberculosis masks always available	-5.5	17.4
Correctly sized tuberculosis masks always available	68.7	26.3
Rent private accommodation	-12.9	24.4
Provided with subsidised doctors quarters on hospital premises	12.9	24.4

Table 4: Hierarchical Bayes Estimation of Average Importances

Attributes	**Average Importances**	**Standard Deviation**
Supervision	9.0	7.2
Practical Experience	22.2	15.4
Resources	15.5	9.1
Rural Allowance	10.6	5.3
Hospital Safety	18.2	14.7
Occupational Hazard	19.1	7.9
Housing	5.5	5.7

The results were then analysed by segments according to gender, future career intentions, and prior undergraduate rural medical exposure. Both males and females valued hospital safety as their highest weighted attribute, and this was more so for females. Female medical students were sensitive to 20% increase in rural allowance. Males more valued being supervised by a consultant, having access to advanced practical experience and fitted TB masks.

Table 5: HB Results Segmented by Gender

	Utility			
Attribute	**Total**	**Male**	**Female**	**Other**
Supervised by Medical Officer	-12.6	-22.0	-9.0	9.4
Supervised by Registrar	-11.0	-10.4	-11.1	-15.5
Supervised by Consultant	23.6	32.3	20.0	6.2
Limited to filling out forms and taking bloods	-76.3	-82.0	-73.9	-70.0
Includes filling out forms, taking bloods and doing procedures e.g., lumbar punctures	76.3	82.0	73.9	70.0
Daily stock out of gloves, syringes and suture packs	-52.8	-51.2	-53.5	-56.3
Gloves, syringes and suture packs available daily	52.8	51.2	53.5	56.3
R4000 (current level)	-36.9	-34.9	-37.3	-51.1
R4340 (8% increase)	15.0	19.5	12.6	17.5
R4800 (20% increase)	22.0	15.4	24.7	33.6
There have been few reports of theft, hijacking and protests in and around the hospital in the past year	61.0	44.5	69.7	41.8
There is a high level of crime in and around the hospital with many reports of theft, hijacking and protests in the past year	-61.0	-44.5	-69.7	-41.8
No tuberculosis masks available in the hospital	-63.2	-67.8	-61.0	-63.7
Poorly fitting tuberculosis masks always available	-5.5	-2.9	-7.0	-0.3
Correctly sized tuberculosis masks always available	68.8	70.7	68.0	64.0
Rent private accommodation	-12.9	-13.6	-12.0	-26.9
Provided with subsidised doctors quarters on hospital premises	12.9	13.6	12.0	26.9

	Importance			
Attribute	**Total**	**Male**	**Female**	**Other**
Supervision	9.0	10.8	8.0	10.1
Practical Experience	22.1	23.5	21.6	20.0
Resources	15.5	15.4	15.5	16.1
Rural Allowance	10.6	10.3	10.7	13.0
Hospital Safety	18.2	14.2	20.4	13.9
Occupational Hazard	19.1	20.1	18.6	18.3
Housing	5.5	5.9	5.2	8.7

(Table 6) Those who intended to specialise gained more utility from being supervised by a consultant and gaining practical experience. Those who intended to join general practice (GP) valued hospital safety and the provision of basic resources higher. (Table 7) Medical students with rural medicine exposure valued hospital safety more highly. Medical students without undergraduate rural medicine exposure preferred basic resources and housing provided more highly.

Table 6: HB Results by Career Intention

	Utility				
Attribute	**Total**	**GP**	**Specialisation**	**Undecided**	**Other**
Supervised by Medical Officer	-12.6	-12.0	-17.6	-6.2	-1.4
Supervised by Registrar	-11.0	-9.5	-10.6	-12.6	-3.9
Supervised by Consultant	23.6	21.4	28.2	18.7	5.4
Limited to filling out forms and taking bloods	-76.3	-65.4	-88.3	-59.8	-71.4
Includes filling out forms, taking bloods and doing procedures e.g., lumbar punctures	76.3	65.4	88.3	59.8	71.4
Daily stock out of gloves, syringes and suture packs	-52.8	-57.0	-48.0	-58.3	-65.7
Gloves, syringes and suture packs available daily	52.8	57.0	48.0	58.3	65.7
R4000 (current level)	-36.9	-37.4	-36.3	-37.5	-40.5
R4340 (8% increase)	15.0	14.5	13.5	16.3	23.7
R4800 (20% increase)	22.0	22.9	22.8	21.2	16.8
There have been few reports of theft, hijacking and protests in and around the hospital in the past year	61.0	70.1	51.9	72.2	73.7
There is a high level of crime in and around the hospital with many reports of theft, hijacking and protests in the past year	-61.0	-70.1	-51.9	-72.2	-73.7
No tuberculosis masks available in the hospital	-63.2	-65.3	-66.5	-59.1	-51.6
Poorly fitting tuberculosis masks always available	-5.5	-15.4	-3.5	-7.1	-6.3
Correctly sized tuberculosis masks always available	68.8	80.7	70.0	66.3	57.9
Rent private accommodation	-12.9	-16.1	-13.4	-12.1	-9.2
Provided with subsidised doctors quarters on hospital premises	12.9	16.1	13.4	12.1	9.2

	Importance				
Attribute	**Total**	**GP**	**Specialisation**	**Undecided**	**Other**
Supervision	9.0	7.3	9.4	8.5	9.2
Practical Experience	22.2	18.7	25.6	17.5	20.4
Resources	15.5	16.3	14.0	17.4	18.8
Rural Allowance	10.6	10.4	10.1	11.5	10.4
Hospital Safety	18.2	20.3	15.5	21.9	21.1
Occupational Hazard	19.1	21.1	19.8	18.1	15.6
Housing	5.5	5.9	5.7	5.2	4.5

Table 7: HB results by Prior Rural Medicine Exposure

	Utility		
Attribute	**Total**	**Rural Med Exposure**	**No Rural Med Exposure**
Supervised by Medical Officer	-12.6	-15.0	-9.3
Supervised by Registrar	-11.0	-10.3	-11.9
Supervised by Consultant	23.6	25.3	21.2
Limited to filling out forms and taking bloods	-76.3	-77.8	-74.4
Includes filling out forms, taking bloods and doing procedures e.g., lumbar punctures	76.3	77.8	74.4
Daily stock out of gloves, syringes and suture packs	-52.8	-49.9	-56.8
Gloves, syringes and suture packs available daily	52.8	49.9	56.8
R4000 (current level)	-36.9	-38.2	-35.2
R4340 (8% increase)	15.0	14.1	16.1
R4800 (20% increase)	22.0	24.2	19.0
There have been few reports of theft, hijacking and protests in and around the hospital in the past year	61.0	65.5	54.8
There is a high level of crime in and around the hospital with many reports of theft, hijacking and protests in the past year	-61.0	-65.5	-54.8
No tuberculosis masks available in the hospital	-63.2	-61.1	-66.1
Poorly fitting tuberculosis masks always available	-5.5	-6.0	-4.9
Correctly sized tuberculosis masks always available	68.8	67.1	71.0
Rent private accommodation	-12.9	-10.3	-16.5
Provided with subsidised doctors quarters on hospital premises	12.9	10.3	16.5

	Importance		
Attribute	Total	Rural Med Exposure	No Rural Med Exposure
Supervision	9.0	8.9	9.1
Practical experience	22.2	22.5	21.7
Resources	15.5	14.7	16.6
Rural allowance	10.6	10.9	10.2
Hospital Safety	18.2	19.2	16.8
Occupational Hazard	19.1	18.6	19.8
Housing	5.5	5.3	5.8

SIMULATIONS

Figure 1 shows a simulation run to predict the share of preference between safe and unsafe hospitals when all other attributes were held to be identical, namely that they would be supervised by a Medical Officer, baseline rural allowance R4000, had access to advanced practical experience, experienced daily stock-outs of basic resources, had no tuberculosis facemasks available, and would rent private accommodation. In that simulation, the safe hospital had 82% of share while the unsafe hospital had 18%. In the next simulation, Figure 2, rural allowance was increased for the unsafe hospital from R4000 to R4800 which resulted in an increased in share of preference from 18% to 34.6% demonstrating that, although influential, increasing rural allowance alone will not overcome the influence of hospital safety on internship job uptake probability.

Figure 1: Simulation Safe vs. Unsafe Hospital

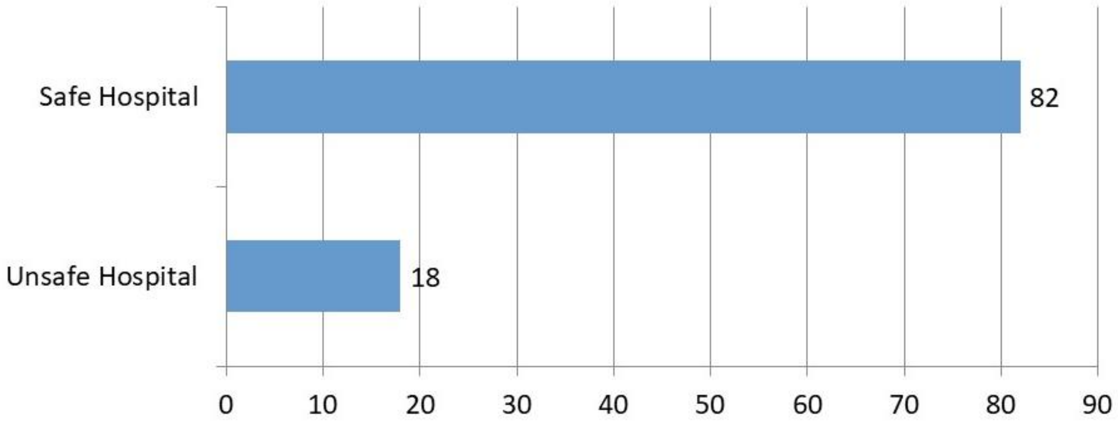
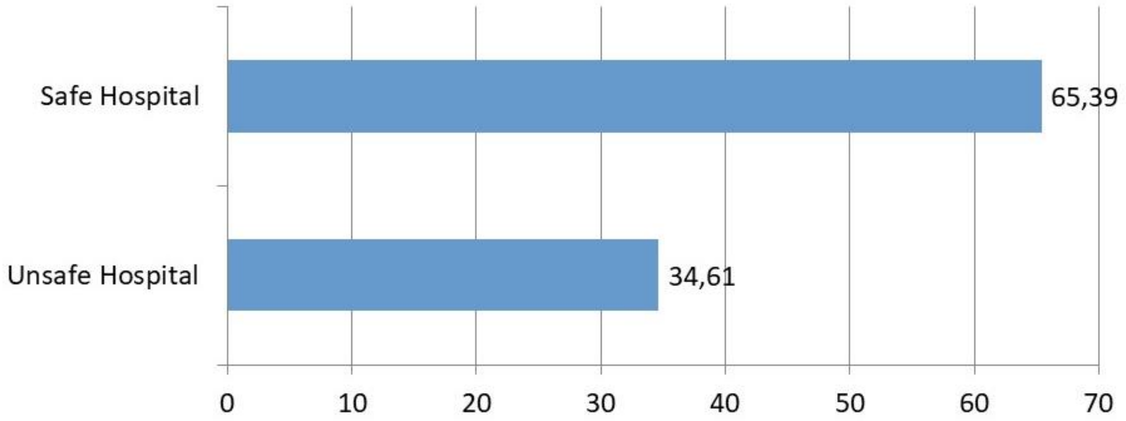


Figure 2: Simulation of Safe Hospital vs. Unsafe Hospital with 20% Rural Allowance Raise



Similarly in Figure 3, a simulation was done to predict the share of preference between hospitals with correctly sized tuberculosis masks and those without when all other attributes were held to be identical, namely that they would be supervised by a Medical Officer, baseline rural allowance R4000, had access to advanced practical experience, experienced daily stock-outs of basic resources, were unsafe, and would rent private accommodation. In that simulation, the facility with the correctly fitted mask had 84.6% of share while the unsafe hospital had 15.4%. In the next simulation, Figure 4, rural allowance was increased for the hospital without tuberculosis masks from R4000 to R4800 which resulted in an increase in share of preference from 15.4% to 25.7% demonstrating that, although influential, increasing rural allowance alone will not overcome the influence of lack of correctly fitted tuberculosis masks on internship job uptake probability.

Figure 3: Simulation Facility With Tuberculosis Mask vs. Facility Without Tuberculosis Masks But a 20% Rural Allowance Raise

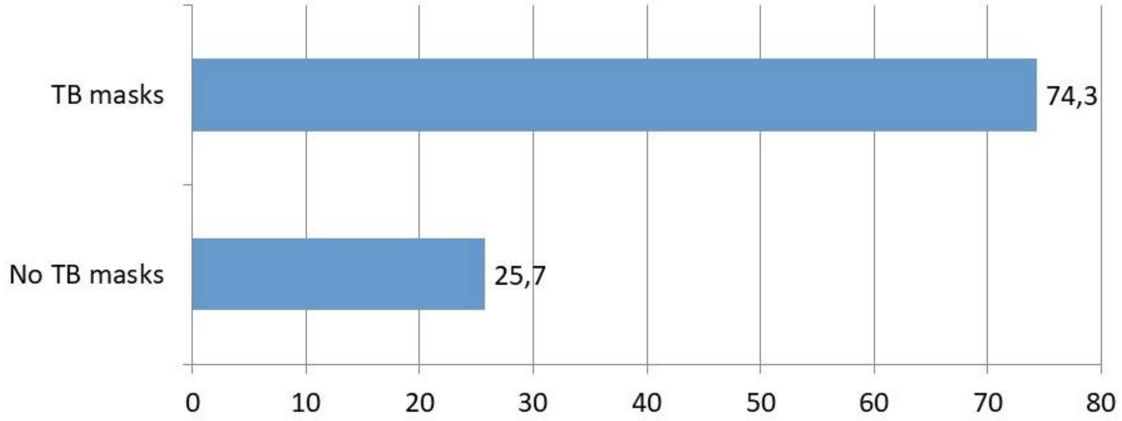
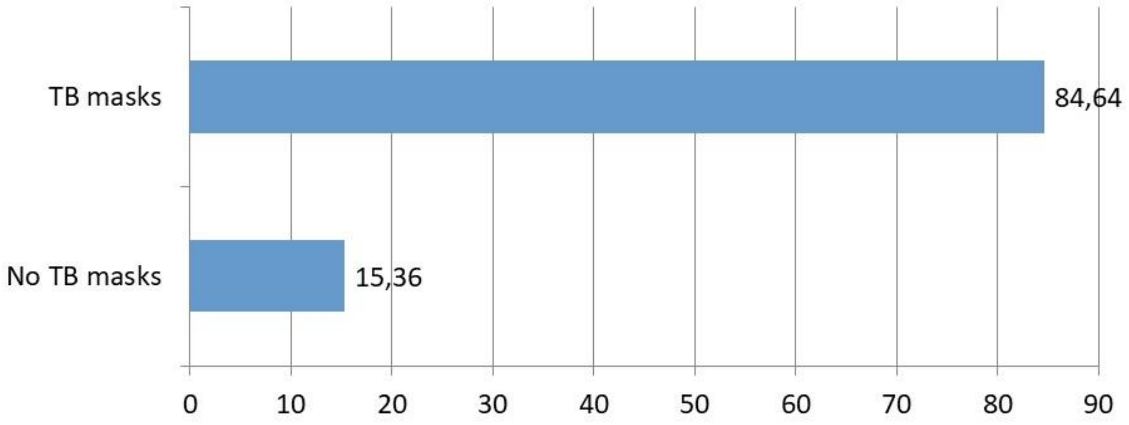


Figure 4: Simulation Facility With Tuberculosis Mask vs. Facility Without Tuberculosis Mask



Limitations of Study

The lack of an opt out option may have led to forced decision-making and is often not considered best practice in marketing studies, but was appropriate to simulate real life situations and make the task more realistic for participants. Only one version of the questionnaire was used and attributes were not randomized, therefore this could contribute to context bias and limit statistical efficiency. Even so, we found only minor reduction in design efficiency for fielding one version compared to more versions. Relative D efficiency for 1 version: 1899.8; Relative D efficiency for 5 versions: 1997.4 = approximately (5% difference).

Conclusion

This study's objective was to explore the heterogeneity in job attribute preferences of final-year medical students for rural health facility internships to inform recruitment policy. The rural allowance should not be a stand-alone incentive for recruitment and retention. Hospital safety, the provision of TB masks and basic resources should be prioritized and may prove to be impactful and cost-efficient in the long-term. To be effective, recruitment packages should take into consideration future career intentions, gender, career aspirations, and previous rural medicine exposure.

Maria Jose

References

Be Safe Paramedical Suppliers. 2019. *Surgical Mask—N95 (20 pack)—Be Safe*. [online] Available at: https://be-safe.co.za/shop/bio-safety/surgical-mask-n95/ [Accessed 6 Mar. 2019].

Burch VC, McKinley D, Van Wyk J, Kiguli-Walube S, Cameron D, Cilliers FJ, Longombe AO, Mkony C, Okoromah C, Otieno-Nyunya B, Morahan PS. 2011. Career intentions of medical students trained in six sub-Saharan African countries. *Education for Health*. **24**:614.

Dambisya YM. A review of non-financial incentives for health worker retention in east and Southern Africa. 2007. *Equinet Discussion Paper*. **44**:49–50.

Lagarde M, Blaauw D. 2014. Pro-social preferences and self-selection into jobs: Evidence from South African nurses. *Journal of Economic Behavior & Organization*. **107**:136–52.

MedicalBrief, June 12th, 2019. https://www.medicalbrief.co.za/archives/attacks-doctors-highlight-security-dangers-state-hospitals/

World Health Organization. 2010. *Increasing access to health workers in remote and rural areas through improved retention: global policy recommendations*. [online] Available at: http://whqlibdoc.who.int/publications/2010/9789241564014_eng.pdf. [Accessed 11 Feb 2019].

Leadership Qualities:
Preferences From the Millennial Generation

Ronald Mellado Miller
Utah Valley University

Christina A. Hubner
Sawtooth Software

Cray Daniel Rawlings
Maureen Andrade
Utah Valley University

Abstract

Utilizing characteristics gathered from business researchers and evolutionary psychology, business students were surveyed to see which characteristics they would prefer to have in a CEO. Using MaxDiff methodology, respondents marked the CEO characteristics that were most and least important to them when considering a future job. Latent Class Multinomial Logit analysis found two distinct groups, a "Sensitive Group" and an "Achievement Group," that preferred contrasting traits in a CEO. Gender differences were also found when investigating a one item TURF analysis using Average Probability Weighted Reach. The study suggests that previous research on preferences in leadership may need modification for the Millennial demographic. Honing in on this unique and upcoming generation can help businesses and other employers understand how to best attract potential candidates by emphasizing particular CEO characteristics as well as hiring CEOs that appeal to their future workforce.

Introduction

The U.S. Bureau of Labor Statistics predicted that by 2015 Millennials would overtake the majority representation of the workforce and by 2030 they will make up seventy-five percent of the workforce (Mitchell, 2015). As more and more college graduates from Generations Y (Millennial) and Z enter the workforce, it is important to understand what these generations prefer in job satisfaction and leadership. Unlike previous generations, Millennials have been shown to be less materialistic and more devoted to global rights and environmental causes. However, they are also known for changing jobs more frequently than those of previous generations (Yeaton, 2008; U.S. Department of Labor Statistics, 2013). While Generation Z aspires to security and stability, as opposed to Millennials' preference for freedom and flexibility, the former shares an openness to switching careers with its generational predecessor (Fourhooks, 2015). Generally, an employer's failure to adhere to these generations' values can lead to job dissatisfaction. Studies indicate that an employee's intention to leave is negatively associated with the employee's perception of the manager's leadership style (Maier, 2011). Thus, in both cases, loyalty to a particular organization is uncharacteristic.

If companies are going to attract their preferred job candidate in a competitive market and keep employees satisfied and productive, they need to understand the preferences of these generations. Millennial turnover costs the U.S. economy an estimated $30.5 billion annually (Fry, 2018). Additionally, The Conference Board surveyed 1,500 individuals on 23 aspects of job satisfaction and found that only fifty-one percent of employees are satisfied with their jobs. Given this context, this study seeks to shed light on the characteristics Millennial- and Generation Z-age cohorts look for in a CEO. Past research has theorized that employees prefer CEOs or leaders that a give them power, money, and/or prestige, and those, who, essentially, seem to have evolutionarily validated characteristics such as physical attractiveness, height, and presence (see Conroy-Beam, 2015 and Pfeffer, 2015). Given that there is no correlation between intelligence and appearance above normal attractiveness levels, these characteristics may lack real utility in the modern world (Pincott 2012). Although research has explored the leadership preferences of Millennials, and to a lesser extent, Generation Z, definitive conclusions are elusive.

Generation Theory

Generation theory, formally known as the Strauss-Howe Generational Theory (1991), based on the work of Karl Mannheim (1952), argues that people share commonalities in values, behaviors, and attitudes based on the social, economic, and historical events that occur during their formative years. Although research (as opposed to anecdotal suppositions) has not unequivocally established differences among generations or come to a consensus on generational characteristics (Robbins & Judge, 2017), generation theory continues to be appealing as a way to explain people's differing perspectives and behaviors.

Many of the Millennial generation, born from the early 1980s to the mid-1990s, have graduated from college and are in the workforce while those born toward the end of this generation will soon begin their professional careers. Current college-age students are predominantly from Generation Z, being born from 1995-2012 (Berlinksky-Schine, 2019); however, there is overlap between generations and some experts extend the Millennial generation timeframe into the year 2000 (Fourhooks, 2015). Also, most universities have significant enrollments of non-traditional students, defined as those 25 years of age or older, thus students in higher education institutions today represent both Millennials and Generation Z.

The U.S. labor force consists of thirty-five percent Millennials and five percent Generation Z (Fry, 2018). Early Millennials are moving into managerial positions while the tail-end of the Millennial generation and Generation Z are preparing for the workforce. Generational theory can extend understanding of the diversity represented by employees in today's organizations and provide insights into how employees can work together effectively. As such, a greater understanding of leadership expectations of today's college students, the workforce of the future, is valuable. A brief overview of the characteristics of Generations Y and Z (Table 1) provides context for this study.

Table 1: Generational Comparison

Generation	Characteristic	Sources
Millennial (Generation Y)	High self-esteem; self-absorbed; self-reliant; autonomous; freedom & flexibility; entitlement and rewards; expect and desire change; frequent job changes; value communication and feedback; collaborative; want meaningful work; results driven; tolerant; technologically savvy; satisfied with companies, jobs, job security, recognition, career development, advancement	Ahmad, 2018; Alexander & Sysko, 2011; Berlinsky-Schine, 2019; Kowske, Rasch, & Wiley, 2010; Laird, Harvey, & Lancaster, 2015; Lowe, 2011; Martin, 2005; Ng, Schweitzer, & Lyons, 2010; Williams & Page, 2011
Generation Z	Anticipate the need for hard work; confident about the future; are online constantly; use multiple technology screens; short attention span; entrepreneurial; high salary expectations; believe science and technology can solve the world's problems; want jobs that will impact the world; wish hobbies could be jobs; value experiences over products; value individuality; independent; competitive; imaginative; responsible; traditional beliefs; value family; self-controlled	Ahmad, 2018; Berlinksy-Schine, 2019; Williams & Page, 2011

Leadership Theory

Early leadership theory posited that leaders were born with certain characteristics that were predictive of leadership (Bass, 1990; Jago, 1982). However, when research failed to identify a set of common traits across those considered to be effective or noteworthy leaders (Mann, 1959; Stogdill, 1948, 1974), the focus turned to leadership behaviors and the idea that individuals can learn to be leaders. That being said, recent research has identified a few characteristics that predict the likelihood of leadership, though not the effectiveness of leadership. These include extroversion, conscientiousness, and openness to experience (Judge, Bono, Illies, & Gerhardt, 2002). Other characteristics supported by research and associated with leadership are confidence, self-efficacy (Smith & Forti, 1998) and emotional intelligence (Goleman, 1995; Mayer, Salovey, & Caruso, 2000).

The most noted of behavioral leadership studies, from Ohio State University (Stodgill, 1948, 1963, 1974) and the University of Michigan (Bowers & Seashore, 1966; Cartwright & Zander, 1960; Likert, 1961, 1967), shared similar results, specifically that leadership involved the technical aspects of work, such as the ability to define roles, structure tasks, and organize people, and people-related aspects, including interpersonal relationships, trust, and respect. These studies are considered follower studies in that they focus on the views of employees. Followership research seeks to explore the expectations and beliefs that people have about leaders (Sy, 2010), relationships between leaders and followers (Bligh, Kohles, & Pillai, 2011), and the roles of leaders and followers (Howell & Mendez, 2008). The current study can be categorized as a trait-based leadership and leadership follower study.

The current study primarily analyzes current leadership trends discussed in “The CEO Next Door: The Four Behaviors That Transform Ordinary People into World-Class Leaders” by Elena L. Botelho and Kim R. Powell from the leadership advisory firm ghSMART; and “Leadership BS: Fixing Workplaces and Careers One Truth at a Time,” by Jeffrey Pfeffer from the Stanford Graduate School of Business. Both seek to answer the controversial question, “What makes a great CEO?” From the ghSMART dataset of over 17,000 leadership assessments, one of the most comprehensive leadership datasets, the former finds that successful CEOs should be decisive, engaging, reliable, and adaptable (Botelho, 2018). Pfeffer, on the other hand, takes a more practical approach to the question. Based on his research he challenges commonly accepted leadership traits, asserting that, “The pursuit of individual self-interest just might be, as virtually all economics writing and theory since the time of Adam Smith teaches, good not just for you but also generally beneficial for the social systems including the work organizations in which you live” (Pfeffer, 2015). With a myriad of theories surrounding effective leadership, it becomes difficult to parse out which ideas hold merit in the workforce.

Generational Leadership Preferences

Identifying generational preferences for leadership traits or behaviors is critical to understanding how to lead, and by extension, to organizational success (McCrindle, 2006; Sujansky, 2004). Preferred traits typically reflect generational values. Millennials have positive attitudes toward work, desire strong workplace relationships, are motivated by challenge, value on-going feedback and training, want advancement opportunities, and seek work-life balance (Eisner, 2005; McCrindle, 2005; Salahuddin, 2010). Both Millennials and Generation Z desire to be meaningfully engaged and have an impact (see Table 1). These characteristics may suggest a preference for transformational leaders who engage and inspire followers to develop their potential (Bass, 1990; Frifth, 2017; Horeczy et al., n.d.). Others argue that Millennials are positioned to be servant leaders, given their commitment to delegation, community, and shared responsibility in order to “learn from and grow with others to be challenged by meaningful work that matches the strengths of the person to their job, and to share and experience life together in accomplishing results” (Balda & Mora, 2011, p. 22).

Millennials like authority, structure, and strong leadership; leaders who unify and take collective action; and those who create change (Society of Human Resource Management, 2004; Zemke, Raines, & Filipczak, 2000). Others claim that Millennials dislike structure (Hewlett, Sherbin, & Sumberg, 2009), preferring less hierarchy and a greater focus on relationships, communication, and creative thinking (Altizer, 2010; Herlett et al., 2009). They place importance on flexibility, empowerment, praise, recognition, feedback, information sharing, involvement, dialogue, project-focused work, clear expectations, and mentoring (Human Resource Professional Association, 2016; McGonagill & Pruyn, 2010; U.S. Department of Commerce, 2011; Verret, 2000). They respect leaders characterized by good work ethic, accountability, strategy, growth, competency, and organizational commitment as well competence, inspiration, loyalty, and determination (Arsenault, 2004; Horeczy et al., n.d.; Salahuddin, 2010). Being caring, imaginative, and ambitious are also highly ranked by Millennials (Salahuddin, 2010). They highly value relationships, mentoring, opportunities for growth, and expect their ideas to be respected; when interest wanes, they will seek employment elsewhere (Dulin, 2008).

As Generation Z has not yet entered the workforce in substantial numbers, less is known about its leadership preferences compared to previous generations although some expect these to be similar to Generation Y (Al-Asfour & Lettau, 2014; United Nations Joint Staff Pension Fund, n.d.). According to some, and similar to the Millennial generation, Generation Z wants transformational leaders who inspire and engage, specifically those with emotional intelligence who share information, openly communicate, and make connections (Frifth, 2017; McCrindle, 2019). They prefer technology-based rather than face-to-face communication; social responsibility and purpose are important, and similar to Millennials, they want performance feedback and mentoring (Zaleski, 2019). Inclusivity (focus on the collective good and collaboration as opposed to competition), curiosity (seek continuing knowledge development, challenge, and world understanding), self-motivation (value freedom and responsibility to perform tasks; flexibility in hours and remote work arrangements), generosity (believe in companies that give back and help others), perseverance (understand the value of hard work and are interested in the struggles of leaders) are valued as part of a mentoring relationship (Patel, 2017).

Although these various listings of generational preferences are helpful and have practical implications for leadership styles and behaviors, it is difficult to determine which of these traits and behaviors are preferred over others and for specific demographic groups. The proliferation of findings present some difficulty in determining just how to respond to generational leadership preferences. The current study is designed to address this gap.

Methods

Participants were gathered using convenience sampling from the largest university in Utah in order to gather data on a substantial number of those about to enter the workforce. Students were given the option to take the survey as part of their courses. A total of 225 respondents of approximately the Millennial age undertook this research programme. 170 of the participants were self-selected men and 56 were self-selected women. Half of the respondents were juniors while 27% of the sample were sophomores, 17% were seniors, and 6% were freshmen. While the vast majority of the sample were Caucasian, a few other ethnicities were sampled as well. 4% of the sample was Hispanic/Latino, 7% Asian, and 3% other.

A MaxDiff survey was constructed using 55 characteristics collected from Botelho, et al., (2018), Conroy-Beam (2015), and Pfeffer (2015), whose combined works include scientific articles, books, a Harvard Business Review article, and, in one case (Pfeffer) from experience leading Stanford's MBA program. Some of the characteristics include appearance, authenticity, charisma, provides status, confidant, good coach, results-oriented, and the like. The characteristics themselves were displayed on a computer and students marked the characteristics that were most and least desirable to them in a CEO. Specifically, the wording utilized was, “Imagine that today you were choosing the ideal CEO you would like to work for. Considering only these attributes, which is the Most Important and which is the Least Important for this decision?” The MaxDiff methodology was helpful in this case due to the long list of items. “Humans are much better at judging items at extremes than in discriminating among items of middling importance or preference” (Louviere, 1993).

Results

Using the hierarchical Bayes Multinomial Logit (HB-MNL) in Sawtooth Software’s Lighthouse Studio program, utilities for each respondent were estimated and later brought into the MaxDiff Analyzer tool. Looking at the sample as a whole, the item “Trustworthy” (Utility = 4.26, Upper CI =4.34 , Lower CI = 4.19) was the most preferred trait in a CEO based on the highest average utility. In line with the preference for a trustworthy CEO, the business students also had higher average utilities for a CEO that was honest, had good moral character, and ethical. Items related to planning and success were also preferred by the sample such as being a hard worker, intelligent, and “Provides the company with growth opportunities.” The least preferred CEO trait was “Stylish appearance” (Utility = 0.01, Upper CI = .02, Lower CI = 0.01). Having a CEO that was religious, politically conservative/liberal, and good looking were not as preferred as other items.

A Latent Class Multinomial Logit analysis was also conducted. A two-group solution was found to be the best fit for the data based on the fit statistics (Percent Certainty = 37.36, AIC = 25621.5, CAIC = 26541.7, BIC = 26432.7, ABIC = 26086.3, Chi- Square = 151543.3, Relative Chi-Square = 139.03). The two-group solution also resulted in a more easily categorized set of groups. Group one was named the “Sensitive Group” (N=126) because they liked CEOs that were supportive, understanding, considerate, authentic, and kind. Group two (N=99), on the other hand, were more achievement focused because they preferred a CEO that was results-oriented, increased the respondent’s income, and provided the company with status.

Using Total Unduplicated Reach and Frequency (TURF) analysis, preferences across respondents and for specific demographic groupings were investigated. The percent reached can reveal how much many of the business students preferred at least one of the items in the bundle. Average Probability Weighted Reach showed that the highest possible reach with one item was 88.11 % (Trustworthy) for the whole sample. The probability of being reached increases to 95.20 % when another top item, such as “Provides the company with growth opportunities,” is included. After two items, the benefit of including more traits for the hypothetical CEO did not lead to a significant increase in reach. Gender differences were found when investigating a one item TURF analysis using Average Probability Weighted Reach. More women than men preferred a CEO that was empowering and a good coach.

Table 2. Gender Differences in Reach

	% Men Reached	% Women Reached
Empowering	55.98%	65.21%
Good Coach	55.41%	65.10%

Discussion

In his famous novel, Anna Karenina, Leo Tolstoy said, “All happy families are alike; each unhappy family is unhappy in its own way.” The results of this study show the opposite effect for those considering their potential CEO. While most respondents agreed on the characteristics that they disliked in a CEO, demographic segmentation found differences in what these groups preferred. A larger percent of women were reached in the TURF when a CEO was a good coach and empowering. The Latent Class Multinomial Logit analysis also showed hidden groups that had differing preferences. The Sensitive Group and Achievement Group were made up of group members with a similar background, and yet, they still were shown to be significantly dissimilar in their CEO preferences.

The current methods resulted in important findings that indicate the optimal direction for future research. Having so many characteristics to compare created a MaxDiff with more items than your traditional MaxDiff. While incentivized, respondents had quite a long survey to get through to view each item 2.5 times. Most participants likely viewed each characteristic as a generally “good” characteristic to have in a CEO which may explain why there was a low variance in the results. Despite these limitations, a fuller coverage of characteristics allowed researchers to develop a more telling method of finding characteristic utility and TURF reach.

A follow-up to this study might use behaviors that demonstrate the characteristics. When presented with the characteristics themselves, participants are subject to bias that may be due to social desirability, idealism, or trait similarity. Instead, future research will present participants with practical instances of the characteristics being demonstrated. This will help control for bias and allow for more polarization in the survey results. Rather than prioritizing between characteristics such as decisive, a good listener, or clear communicator, the participants would see an example that shows the behavior. For example: “Follows through on actions, promises, and assignments,” “Listens more than speaks,” or “Expresses ideas clearly in written communication.” Future studies thereby might lead to insights into what a good leader should do, rather than what general traits he or she should have.

Conclusion

Alastair Mitchell, co-founder and CEO of Huddle.com, states it best, “a key piece of advice for organizations bracing themselves for the Millennial invasion is: listen and learn” (Mitchell, 2015). The results from the hierarchical Bayes Multinomial Logit and TURF analysis showed noteworthy implications in this Millennial sample study. Among workers, there are two distinct latent classes: Sensitive and Achievement. This has implications for how organizations set objectives and priorities internally. For example, a cover-all approach to motivating employees would inevitably hinder one of the two mutually exclusive groups. Catering to both groups could increase success and efficiency. Important differences were also found between theoretical and practical standpoints. Where Pfeffer feels workers take a more pragmatic view of their leaders; the results show that the view is more complex than he has found. This suggests a need for more research.

The current leaders of this generation are notorious for portraying a “fake,” or idealized, version of themselves through social media and “fake news.” To navigate this world of “fake” news, Millennials will often compare themselves to an ideal or harbor skepticism for news. With this perspective, the study finds that trustworthiness and good moral character are actually some of the highest preferred items. The rising generation desires leadership practices that go against the current norms, and it would be wise to model corporate strategies accordingly.

Overall, Sawtooth Software’s tool was able to not only parse out differences that were important from both theoretical and practical standpoints but also was able to rank them in terms of preference. This utility cannot be overemphasized as most graduates and businesses are currently surrounded in a sea of experts emphasizing different desirable characteristics. This analysis brings clarity to the actual desires of employees likely to be hired in the next round of commerce and advances the use of a statistical tool capable of lending insight to that process.

Ronald Miller

Christina Hubner

Cray Rawlings

Maureen Andrade

References

Ahmad, I. (2018). Comparing the differences between generation Z and millennials. Retrieved from https://www.socialmediatoday.com/news/comparing-the-differences-between-generation-z-and-millennials-infographic/517903/

Al-Asfour, A., & Lettau, L. (2014). Strategies for leadership styles for multi-generational workforce. *Journal of Leadership, Accountability and Ethics 11*(2), 58–69. Retrieved from http://www.na-businesspress.com/JLAE/Al-AsfourA_Web11_2_.pdf

Alexander, C. & Sysko, J. (2011). A study of the cognitive determinants of generation Y's entitlement mentality. *Proceedings of the Allied Academies International Conferences, 17*(1), pp. 1–6. Arden, NC: The DreamCatchers Group LLC. Retrieved from https://www.abacademies.org/Public/Proceedings/Proceedings28/AE%20Proceedings%20Spring%202011.pdf

Altizer, T. E. (2010) Motivating gen Y amidst global economic uncertainty. *Journal of Learning in Higher Education, 6*(1), 44–54. Retrieved from jwpress.com/JLHE/Issues/JLHE-Spring2010.pdf#page_44

Arsenault, P. M. (2004). Validating generational differences: A legitimate diversity and leadership issue. *Leadership & Organization Development Journal, 25*(2), 124–141.

Balda, J. B., & Mora, F. (2011). Adapting leadership theory and practice for the networked, millennial generation. *Journal of Leadership Studies, 5*(3), 13–24.

Bass, B. M. (1990. Bass and Stogdill's handbook of leadership: A survey of theory and research. New York: Free Press.

Berlinsky-Schine, L. (2019). *Millennials in the workforce: 5 positive changes they're making*. Retrieved from https://fairygodboss.com/articles/millennials-in-the-workforce#

Berlinsky-Schine, L. (2019). *Generation Z: 18 statistics about today's newest workers*. Retrieved from https://fairygodboss.com/articles/gen-z-statistics

Bligh, M., Kohles, J., & Pillai, R. (2011). Romancing leadership: past, present and future. *The Leadership Quarterly, 22*(6), 1058–1077.

Botelho, E., Powell, K., & Raz, T. (2018). The Ceo next door the 4 behaviours that transform ordinary people into world-class leaders. London: Virgin Books.

Bowers, D. G., & Seashore, S. E. (1966). Predicting organizational effectiveness with a four-factor theory of leadership. *Administrative Science Quarterly, 11*(2), 238–263.

Cartwright, D., & Zander, A. (1960). *Group dynamics research and theory*. Evanston, IL: Row, Peterson.

Conroy-Beam, D., Buss, D. M., Pham, M. N., & Shackelford, T. K. (2015). How Sexually Dimorphic Are Human Mate Preferences? *Personality and Social Psychology Bulletin*, 41(8), 1082–1093. doi: 10.1177/0146167215590987

Dulin, L. (2008). Leadership preferences of a generation Y cohort: A mixed-methods investigation. *Journal of Leadership Studies, 2*(1), 43–59. doi:10.1002/jls.20045

Eisner, S. E. (2005). Managing generation Y. *Society for the Advancement of Management,70*(4), 4–15.

Fourhooks. (2015, April 26). *The generation guide—Millennials, gen X, Y, Z and baby boomers*. Retrieved from http://fourhooks.com/marketing/the-generation-guide-millennials-gen-x-y-z-and-baby-boomers-art5910718593/

Frith, B. (2017, September 4). *Millennial workers prefer transformational leaders*. Retrieved from https://hrmagazine.co.uk/article-details/millennial-workers-prefer-transformational-leaders

Fry, R. (2018, April 11). *Millennials are the largest generation in the U.S. labor force*. Retrieved from https://www.pewresearch.org/fact-tank/2018/04/11/millennials-largest-generation-us-labor-force/

Goleman, I. R. (1995). *Emotional intelligence*. New York: Bantam.

Hewlett, S., Sherbin, L., & Sumberg, K. (2009). How gen Y & boomers will reshape your agenda. *Harvard Business Review, 87*(7/8), 71–76.

Horeczy, A., Lalani, Al, Mendes, G., Miller, M., Samsa, L., & Scongack, T. (n. d.). *Leadership preferences of generation Y*. Retrieved from http://seanlyons.ca/wp-content/uploads/2012/01/Leadership-Preferences-of-Gen-Y.pdf

Howell, J., & Mendez, M. (2008). Three perspectives on followership. In R. Riggio R, I. Chaleff, & J. Lipman-Blument (Eds.), *The art of followership: How great followers create great leaders and organizations* (pp. 25–39). San Francisco: Jossey-Bass.

Human Resources Professionals Association. (2016). *HR & millennials: Insights into your new human capital*. Retrieved from https://www.hrpa.ca/Documents/Public/Thought-Leadership/HRPA-Millennials-Report-20161122.pdf

Jago, A. G. (1982). Leaderships: Perspectives in theory and research. *Management Science, 28*(3), 315–366.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780.

Kowske, B. J., Rasch, R., & Wiley, J., (2010). Millennials' (lack of) attitude problem: An empirical examination of generational effects on work attitudes. *Journal of Business and Psychology*, *25*(2), 265–279. Retrieved from https://doi.org/10.1007/s10869-010-9171-8

Laird, M. D., Harvey, P., & Lancaster, J. (2015), Accountability, entitlement, tenure, and satisfaction in Generation Y. *Journal of Managerial Psychology, 30*(1), 87–100. Retrieved from https://doi.org/10.1108/JMP-08-2014-0227

Likert, R. (1961). *New patterns of management*. New York: McGraw-Hill.

Likert, R. (1967). The human organization: Its management and value. New York: McGraw-Hill.

Mann, R. D. (1959). A review of the relationship between personality and performance in small groups. *Psychological Bulletin, 56*(4), 241–270.

Mannheim, K. (1952). The problem of generations. In P. Kecskemeti (Ed.), *Essays on the sociology of knowledge: Collected works*, *Volume 5 (*pp. 276–322). New York: Routledge.

Thomas A. Maier (2011). Hospitality Leadership Implications: Multigenerational Perceptions of Dissatisfaction and Intent to Leave, Journal of Human Resources in Hospitality & Tourism, 10:4, 354–371, DOI: 10.1080/15332845.2011.588503

Mayer, J. D., Salovey, P., & Caruso, D. R. (2000). Models of emotional intelligence. In R. J. Stermberg (Ed.), *Handbook of intelligence* (pp. 396–420). Cambridge: Cambridge University Press.

McCrindle, M. (2006). New generations at work attracting, recruiting, retraining & training generation Y. Baulkham Hills, N.S.W: McCrindle Research.

McCrindle. (2019). *Leadership*. Retrieved from https://generationz.com.au/portfolio/leadership/

McGonagill, G., & Pruyn, P. (2010). *Leadership development in the U.S.: Principles and patterns of best practice*. Retrieved from http://www.mobiusleadership.com/resources/Leadership_Development_in_the_US-_Best_Practice-TD%2020jan10_FINAL.pdf

Alastair Mitchell, H. (2015, August 7). The Rise of the Millennial Workforce. Retrieved from https://www.wired.com/insights/2013/08/the-rise-of-the-millennial-workforce/

Ng, E., Schweitzer, L., & Lyons, S. (2010). New generation, great expectations: A field study of the millennial generation. *Journal of Business and Psychology, 25*(2), 281–292.

Patel, D. (2017, August 27). *The top 5 traits gen z looks for in leaders*. Retrieved from https://www.forbes.com/sites/deeppatel/2017/08/27/the-top-5-traits-gen-z-looks-for-in-leaders/#53b3c4f9609d

Pfeffer, J. (2015). Leadership Bs: fixing workplaces and careers one truth at a time. New York: Harper Business.

Pincott, J. E. (2012, November 5). What Your Face Really Reveals About You. Retrieved November 22, 2019, from https://www.psychologytoday.com/us/articles/201211/what-your-face-really-reveals-about-you.

Robbins, S. P. & Judge, T. A. (2017), *Organizational Behavior*. (17th ed.). Pearson, New York.

Salahuddin, M. M. (2010). Generational differences impact on leadership style and organizational success. *Journal of Diversity Management, 5*(2), 1–6.

Society for Human Resource Management. (2004). *Leadership styles series part ii: Leadership styles.* Retrieved from http://multigen.shrmindia.org/resources/articles/leadership-styles-series-part-ii-leadership-stylesgenerational-differences

Stogdill, R. M. (1948). Personal factors associated with leadership: a survey of the literature. *Journal of Psychology, 25*(1), 35–71.

Stogdill, R. M. (1963). *Manual for the Leader Behavior Description Questionnaire form XII.* Columbus: Ohio State University, Bureau of Business Research.

Stogdill, R. M. (1974). Handbook of leadership: A survey of theory and research. New York: Free Press.

Strauss, W., & Howe, N. (1991). *Generations: The history of America's future, 1584–2069*. New York: William Morrow and Company.

Sujansky, J. (2004, April). Leading a multi-generational workforce. *Occupational Health and Safety, 73*(4), 16–18.

Sy, T. (2010). What do you think of followers? Examining the content, structure, and consequences of implicit followership theories. *Organizational Behavior and Human Decision Processes, 113*(2), 73–84.

U.S. Department of Labor Statistics. (2013). *Tenure of American workers*. Retrieved from www.bls.gov/spotlight/2013/tenure/pdf/tenure.pdf

United States Department of Commerce. (2011). *Traditionalists, boomers, x'ers and nexters-NOAA's generational diversity at work*. National Oceanic and Atmospheric Administration, U.S. Department of Commerce. Retrieved from http://www.rdc.noaa.gov/~Diversity/genarticle.html Verret, C. (2000). Generation y: Motivating and training a new generation of employees. Retrieved from http://www.hotel-online.com/Trends/CarolVerret/GenerationY_Nov2000.html

United Nations Joint Staff Pension. (n.d.). *Traditionalists, baby boomers, generation x, generation y (and generation z) working together*. United Nations Joint Staff Pension Fund, Retrieved from http://www.un.org/staffdevelopment/pdf/Designing Recruitment, Selection & Talent Management Model tailored to meet UNJSPF's Business Development Needs.pdf

Verret, C. (2000). *Generation y: Motivating and training a new generation of employees*. Retrieved from http://www.hotel-online.com/Trends/CarolVerret/GenerationY_Nov2000.html

William, K., & Page, R. A. (2011). Marketing to the generations. *Journal of Behavioral Studies in Business, 3*(3), 1–17. Retrieved from http://www.aabri.com/manuscripts/10575.pdf

Yeaton, K. (2008). Recruiting and managing the 'why?' generation: Gen Y. *The CPA Journal, 78*(4), 68–72.

Zaleski, A. (2019, February 1). What today's leaders need to know about generation Z. Retrieved from https://www.greenbaypressgazette.com/story/sponsor-story/st-norbert-college/2019/02/01/what-todays-leaders-need-know-generation-z/2614257002/

Zemke, R., Raines, C., & Filipczak, B. (2000). Generations at work: Managing the clash of veterans, boomers, xers, and nexters in your workplace. New York: NY: AMACOM.

Virtual Reality Meets Traditional Research: or the Reality behind Virtual Reality Enhanced Interviews

Alexandra Chirilov
GfK SE, Germany

Introduction

Virtual Reality (VR) is becoming more widely available to consumers. The GfK Tech Trends Report shows increased consumer interest in VR, with 40% of UK consumers, above 18 years old, intending to buy a VR device in the next 12 months.

At the same time, more and more industries are using the power of VR. From carmakers to tourism providers, brands are using VR as a new opportunity to reach new customers and build stronger relationships. Businesses communicate the value of their products through VR with product demos, 360 tours, virtual shopping, dressing rooms, or showrooms.

For marketers, VR has gained some ground already, while marketing researchers are confidently looking out for ways to assess the feasibility and the scalability of VR-enhanced interviews. VR offers a new set of tools to provide a richer, more immersive experience that allows us to test products that are more realistic and to visit the environment in which the customers experience the product(s), while at the same time having full experimental control over features and conditions. This sounds like an incredible technology with endless possibilities for Market Research (MR), from store layout tests to car clinics, to product development and much more. But, can really any person participate and engage in such an experience? Will they act just as they would do in a real-world situation? Perhaps yes, but to what extent must this still be evaluated before we move into a VR market research world with blind faith? How should data from this new research environment be interpreted? How does it relate to the more traditional data collection methods?

For all of the above reasons, GfK decided it was time to go beyond a basic evaluation and conduct a full validation study using a systematic experimental design to deliver strong recommendations on using VR in MR.

Research Objectives

Probably most of the marketing research studies we conduct today inherently have the potential flaw that the act of completing a survey does not mimic the natural decision-making process used by consumers. Definitely, the use of VR gives respondents a more realistic decision-making context compared to the more traditional approaches. However, it is important to understand if and how the benefit of realism we bring into the design impacts the data quality and predictive power. This paper will show how VR can be used in MR and will discuss the benefits of VR-enhanced interviews. The presentation concludes with a discussion of the business implications of using VR as a data collection technique.

In order to compare the traditional research approach—CAPI (Computer Assisted Personal Interview)—with the VR approach, an empirical study was conducted using a sequential monadic design. The study was designed to help us reach the following objectives:

- Evaluate the feasibility of VR as a new method for data collection.
- Compare the user experience in the two study environments (VR vs. CAPI) and the impact on respondents' engagement.
- Understand the effect of different study environments (VR vs. CAPI) on the respondents' product preferences and behavior.
- Compare the in- and out-of-sample predictive power of the two study environments (VR vs. CAPI).

RESEARCH METHODOLOGY

This empirical research utilized both quantitative and qualitative methodologies to examine the effectiveness of VR-enhanced interviews versus CAPI. Both stages of the research were conducted in a central location in Nuremberg, Germany, during January–April 2017. Respondents were recruited from the pool of visitors to a popular indoor event location. The interviews took place in a separate room on the location's premises.

During the quantitative stage of the research, we asked more than 250 respondents untrained in VR to complete similar conjoint tasks in VR, using the HTC Vive headset, and in CAPI, using an iPad (12.9" screen). The order of the two techniques was rotated between respondents to reduce possible position biases. The study participants were screened to be at least 18 years old and own a car or intend to buy one in the next thirty-six (36) months. No other quotas were used (besides ensuring an equal number of participants per rotation). For the second stage of the research, in-depth interviews were conducted with a smaller number of respondents (n=8) right after they finished the quantitative part with the purpose of collecting additional insights about their experience.

For the conjoint exercise, the sample was randomly divided into 15 versions of 9 tasks. In each task, 5 concepts were shown plus the alternative "none." Respondents were asked to select the concept they preferred most in each of the 9 tasks: 7 random tasks and 2 holdouts to test the in-sample predictive power. To minimize the tendency to reproduce the choices from the first experiment into the second experiment, respondents were allocated to different versions (e.g., a respondent was randomly allocated to version number 1 in CAPI and to version number 7 in VR). We also explicitly mentioned in the survey that the two parts are similar but not identical and that each task is unique.

We had two samples of respondents: 1) the main sample of 200 respondents that took the full survey with all 9 tasks, and 2) the benchmark sample of 50 holdout respondents tested one fixed CBC design with five tasks to test the out-of-sample predictivity.

The research focus area of this study was the automotive sector. We wanted to use relatively common products that most consumers can highly relate to and which are highly relevant for our industry, too.

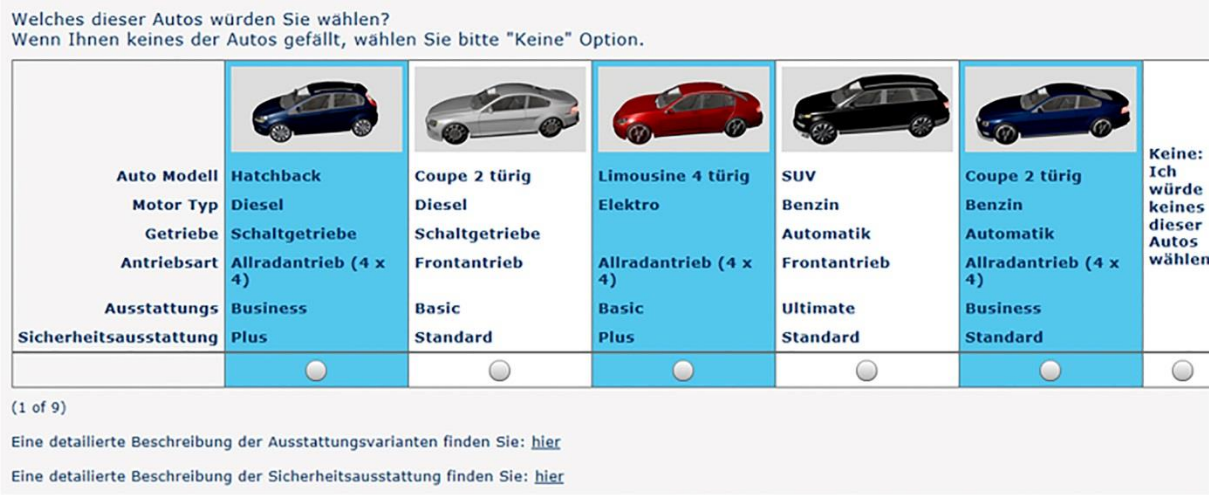
For the conjoint exercise, eight attributes were included: 3 visual attributes and 5 text-based attributes. The visual attributes were car type, colour, and wheel type. The text-based attributes were engine, transmission, drive train, equipment, and safety. Attributes and levels were carefully chosen to ensure their relevance for all respondents. Respondents were asked to assume that all cars belong to the same price category to decrease the complexity of the task and learn about respondent’s preference beyond the price constraints. More specifically, the attributes and levels included in the study are detailed in Table 1. The visual attributes were represented in the CAPI experiment using 2D pictures, while in VR we used 3D models. The experimental design was generated in Sawtooth Lighthouse Studio v9.2 using the balanced overlap generation method.

Table 1: Attributes and Levels

	Attributes	Level 1	Level 2	Level 3	Level 4
Visual	Car Type	SUV	Hatchback	Limousine	Coupé
	Colour	Black	Silver	Red	Blue
	Wheel	Type 1	Type 2	Type 3	
Text	Engine	Petrol	Diesel	Electric	
	Transmission	Automatic	Manual		
	Drive Train	Front-wheel drive	Four-wheel drive		
	Equipment Level	Basic	Business	Ultimate	
	Safety Level	Standard	Plus		

Below is how the actual conjoint tasks appeared in CAPI and in VR (Figure 1 and Figure 2).

Figure 2: Screenshot of One of the CAPI Tasks



Welches dieser Autos würden Sie wählen?
Wenn Ihnen keines der Autos gefällt, wählen Sie bitte "Keine" Option.

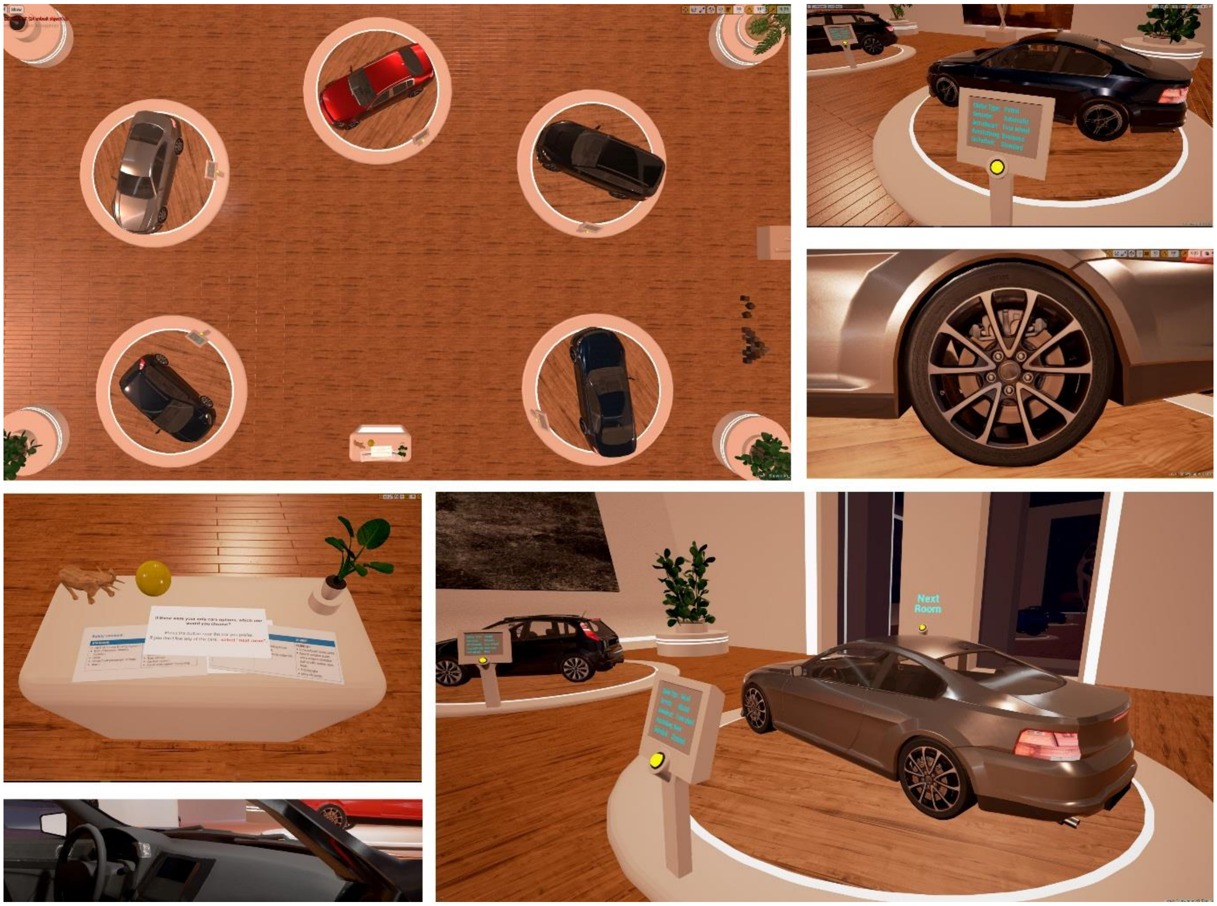
Auto Modell	Hatchback	Coupe 2 türig	Limousine 4 türig	SUV	Coupe 2 türig	Keine: Ich würde keines dieser Autos wählen
Motor Typ	Diesel	Diesel	Elektro	Benzin	Benzin	
Getriebe	Schaltgetriebe	Schaltgetriebe		Automatik	Automatik	
Antriebsart	Allradantrieb (4 x 4)	Frontantrieb	Allradantrieb (4 x 4)	Frontantrieb	Allradantrieb (4 x 4)	
Ausstattungs	Business	Basic	Basic	Ultimate	Business	
Sicherheitsausstattung	Plus	Standard	Plus	Standard	Standard	
	○	○	○	○	○	○

(1 of 9)

Eine detaillierte Beschreibung der Ausstattungsvarianten finden Sie: hier

Eine detaillierte Beschreibung der Sicherheitsausstattung finden Sie: hier

Figure 3: Screenshot of One of the VR Tasks



VR Technology

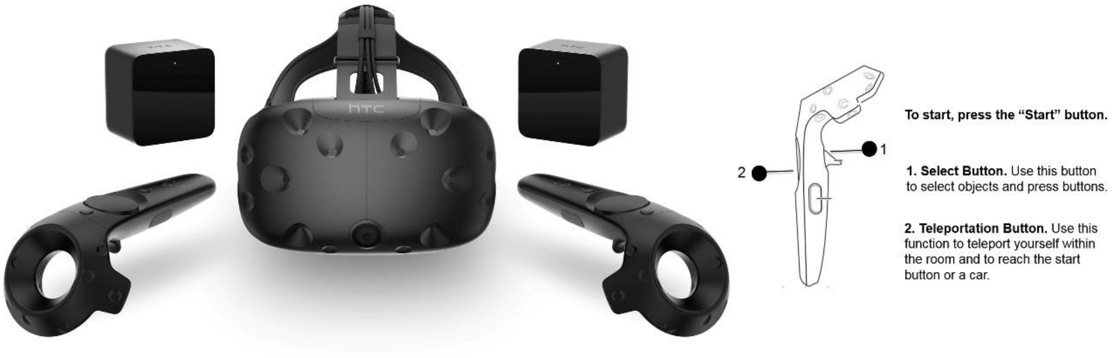
VR is the use of computer technology to create a simulated environment. Unlike traditional user interfaces, VR places the user inside the experience. Instead of looking at a monitor in front of them, users are immersed and able to interact with 3D models. VR facilitates a realistic display of the products' appearance (material, surface, colours) and functionality.

For this research, we used the HTC Vive headset and controllers (Figure 3). The respondents could walk around in a 4.5m by 4.5m tracking space. The HTC Vive is one of the most immersive high-end headsets on the market, which includes a display featuring two 1080 x 1200 screens, one for each eye. This gives the Vive a total resolution of 2160 x 1200 pixels, and an aspect ratio of 9:5 (HTC Vive Technical Specification, 2016).

An operator instructed the respondents on the proper use of the headset and controllers and explained the optimal way to control the application (teleportation and select options). Pretests showed that many of the respondents spent their first few minutes in VR either very energized or hesitant. Therefore, to facilitate the familiarization process, we first placed the respondents into a VR training room. Several interactive elements (ball, Lego cubes, etc.) were placed in there to be used by participants to train their skills in VR and to reduce their enthusiasm and need for play.

During the VR experiment, the researcher was able to observe on a separate screen the respondents' actions (Figure 4). Additionally, the respondents' movements in the VR app were recorded to be further analysed.

Figure 4: HTC Vive Headset and Controllers

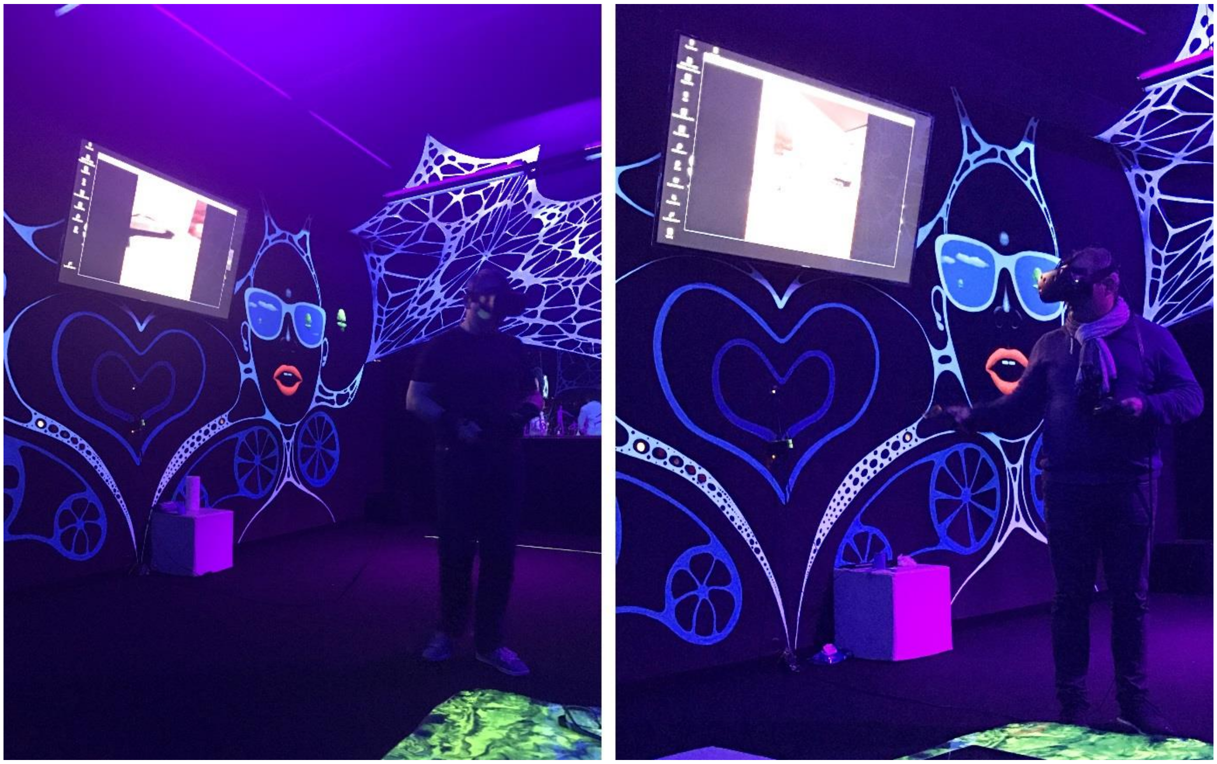


To start, press the "Start" button.

1. Select Button. Use this button to select objects and press buttons.

2. Teleportation Button. Use this function to teleport yourself within the room and to reach the start button or a car.

Figure 5: Respondents Using HTC Vive Glasses to Complete the VR Experiment



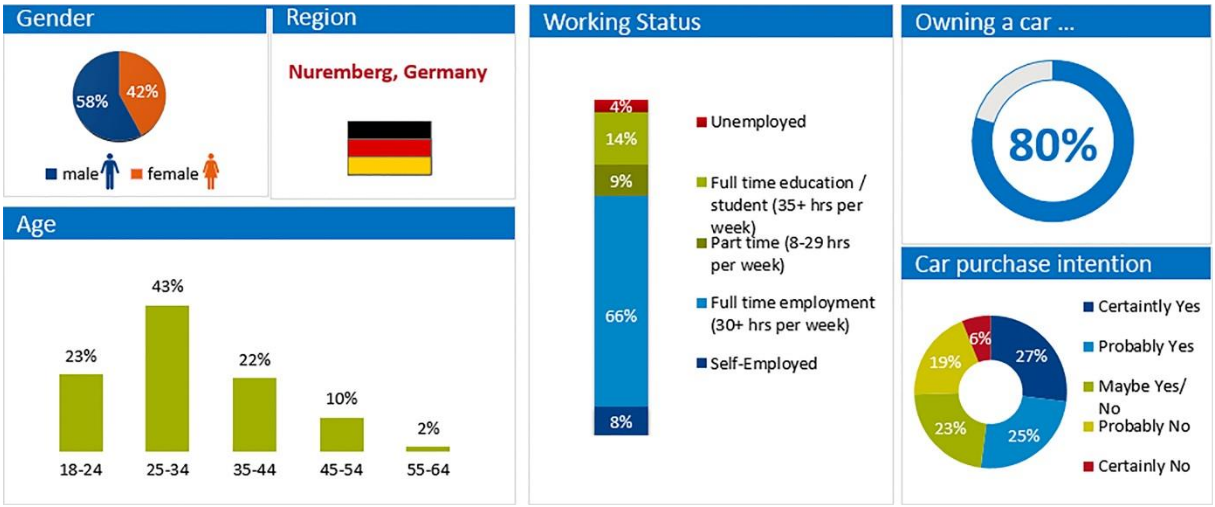
Results

Sample Description

We achieved a good main sample mix (natural fallout, no quotas), as seen in Figure 5. To ensure the comparability between the main sample and benchmark sample, the holdout respondents were weighted using RIM weighting procedure in R (rim-efficiency =72.7%). The target weights were based on the composition of the main sample. The weighting criteria were: age, gender, working status, car ownership, intention to buy a car.

As an additional fieldwork observation, the younger males more proactively asked the interviewer to take part in the VR experiment. Generally, the recruitment process went relatively easy, and most of the people who were asked to participate in the experiment accepted without hesitation (this positive reaction was mainly triggered by the possibility to test VR).

Figure 6: Socio-Demographic Profile of the Respondents



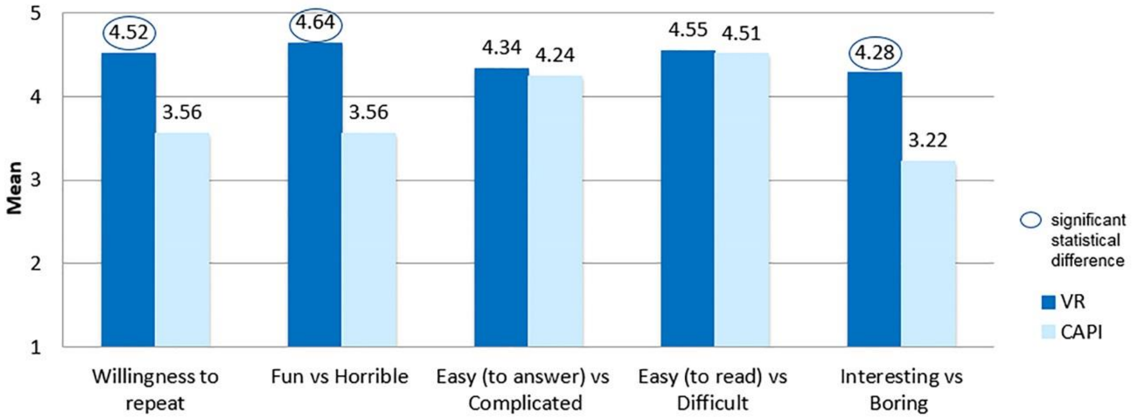
The findings of this research are organized, considering the objectives previously stated.

Respondents' Feedback

Respondents' Satisfaction

The respondents were asked to specifically think about the conjoint exercise they had just completed and answer a set of five questions on a 5-point scale, the higher the mark the more positive the experience was perceived, as seen in Figure 6.

Figure 7: Respondents' Satisfaction



Regardless of the socio-demographic profile or if VR was first in the rotation or not, respondents evaluated the VR interview significantly better than CAPI. For respondents, the experience was far more fun and interesting. Also, their willingness to repeat the VR experience is significantly higher.

Qualitative Assessment

The in-depth interviews were conducted face to face with a trained qualitative moderator from GfK qualitative research group. The moderator asked for feedback about the VR as well as the CAPI experience. What we noticed was that the participants had a strong desire to talk about the VR experience. Therefore, we first provided them with the opportunity to express any opinion about the technology before moving to insights.

When assessing the VR experience, respondents talked about the methodology as being pleasant, very exciting, and very easy to use as well. Some comments from the qualitative research component are as follows:

- "It [VR] was quite pleasant and very exciting for me."
- "The experiment on the tablet was a little bit boring, but it was okay. The virtual reality was more fun even though I was a little dizzy in the end."
- "It was very easy and clear."
- "It feels very natural, the movement comes from the body, the movements have the same directions, and it is very well adjusted. I felt very comfortable with it. It was very easy to use."

Also, some respondents stated that the high similarity to their real-world purchase environment provided a better ability to simulate actual decision-making processes. Below are some of the respondents' comments:

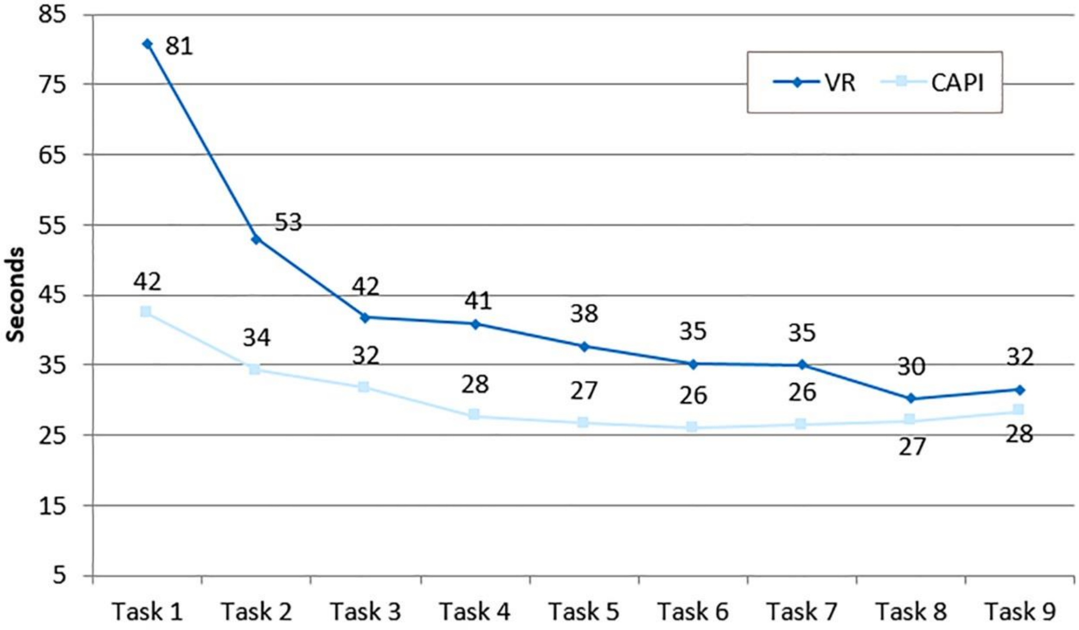
- "On the tablet, it was more difficult to differentiate between the designs of the cars. You can see that much better in virtual reality."

- "It's a big difference from tablet to virtual reality but not from the virtual reality to the real world because you basically do the same thing. You look at the car and see what aspects it has."
- "The VR models are all totally different like it is when you go shopping. On the tablet, I was like 'no, no, no, no,' to all the models. In virtual reality, the models seem to be more interesting."
- "I think what I did in the virtual reality is closer to buying a car in the real world."

Interview Quality

On average, the respondents needed only 40% more time to answer the VR interview compared to the CAPI one (6.43 vs. 4.51 mins). But, the respondents spent as long in the VR training room (6.46 mins) as they spent in the actual VR interview. The learning process continued over the first 3 out of 9 tasks while afterward, the time per task between CAPI and VR tends to level out, as illustrated in Figure 7. Considering that in VR the respondents physically needed more time to read all car characteristics and compare them, as they had to teleport themselves from one car to another, we can conclude that the respondents spent less time in the decision-making process in VR compared to CAPI.

Figure 8: Time per Task



Regardless of the technology type, there are no inherent biases regarding the selection of concepts by position (e.g., always selecting the car placed in the same position in all tasks). However, VR provides an even better quality interview. The bad respondent rate was lower in VR (3.5%) compared to CAPI (4.5%). A bad respondent is a respondent who chose the None option in more than 75% of tasks, or chose the same concept position in more than 75% of tasks, *or* achieved a low RLH (0.25 for 5 concepts).

In both parts of the VR interview (tasks 1–5 vs. tasks 6–9), respondents selected each alternative a similar number of times. In comparison to this, during the second half of the CAPI interview, respondents selected slightly more often one of the first two alternatives (+5.8%), which were placed closer to the "next" button. A possible explanation can be that the engagement and motivation in CAPI decreases faster and so the respondents tried to move quicker through their later tasks.

Additionally, respondents selected the None option less often in VR (11%) than in CAPI (18%). Even though the None alternative was less prominent in VR than in CAPI, the respondents were notified both verbally by the operator and via the written instructions about the possibility to select the None option. Therefore, it is possible that respondents perceived the 3D car models more relevant and appealing; and respectively, that the VR experiment was overall more engaging.

Some more important VR specific highlights:

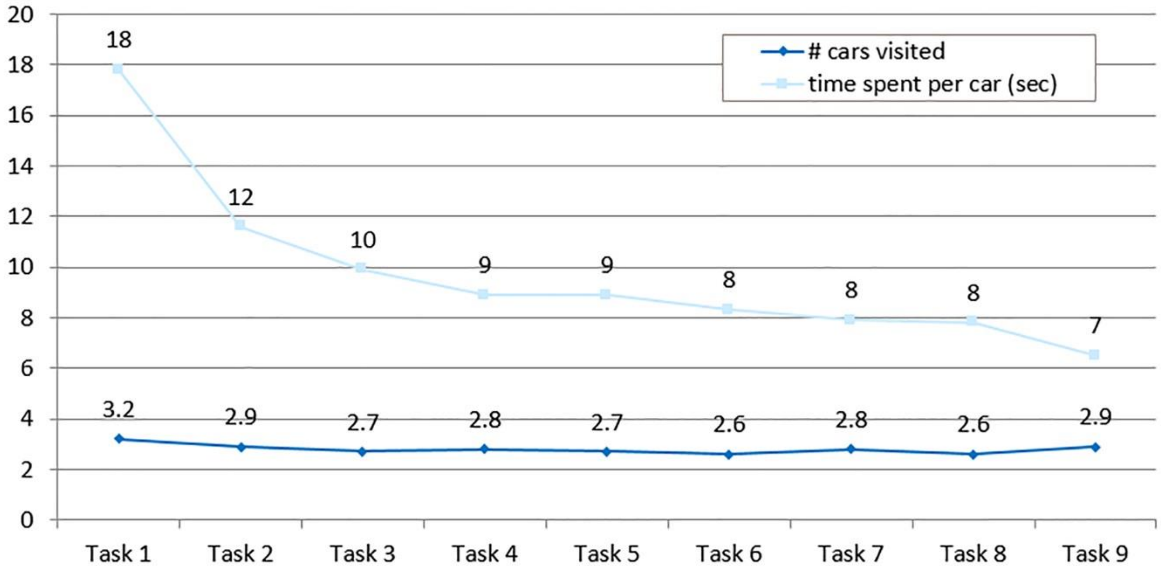
- None of the respondents developed motion sickness.
- Only two respondents dropped out because of the technical difficulties; all other respondents were able to properly use the VR headset and controllers.

Location Data

Additionally, we tracked respondents' movements in the VR conjoint tasks (e.g., movement patterns between cars). Therefore, we can provide additional insights regarding the participants' behaviour in VR.

On average, respondents visited 2.8 cars out of 5 in each task, and they spent 9.7 seconds per car. The number of cars visited per task didn't decrease over tasks but the time spent per car did, as seen in Figure 8. The fact that we saw no decrease in the number of cars visited per task and only saw a decrease in the number of seconds spent per car (otherwise a normal learning effect) proves that this behaviour is not a reaction to fatigue or disengagement but a consistent decision-making strategy across all tasks.

Figure 9: Number of Visited Cars and Time Spent per Car



A visit is defined as the presence of the respondents in the close proximity of a car while having the opportunity to closely inspect the car and read the additional information about the car configuration.

By employing additional analysis, we identified that the vast majority of respondents (84%) simplified their choices:

- In the first stage, without actually visiting the cars, respondents formed a "consideration subset" of cars that matched their visual preferences in regards to the car type, colour, or wheel type (the size of the consideration subset was, on average, 2.4 cars per task).
- In the second stage, respondents made a final choice for one car by visiting and carefully evaluating the cars in the consideration subset. With regards to the selected car, in the same task, respondents visited it on average 1.6 times (vs. 1.1 times for the others in the subset), and they spent on average 3.6 extra seconds evaluating it (12.6 vs. 9.0 seconds for the others). Considering this, we can conclude that respondents involved more complex heuristics in this second stage (final choice).

This two-stage process is well-established in the academic literature as a realistic description of the process by which people make decisions (Payne, 1976; Gaskin, 2007).

The rest of the respondents (16%) visited 4 or even all alternatives in a task before deciding which car they preferred most. However, we are missing evidence regarding their decision-making pattern. It is very likely that they also simplified their decision by using different heuristics that are text-based (about engine type, equipment level, safety level, etc.).

Below we present the path graph of two representative respondents from each of the two groups (Figure 9 and Figure 10).

Figure 10: Path Graph per Respondent (Group 1)

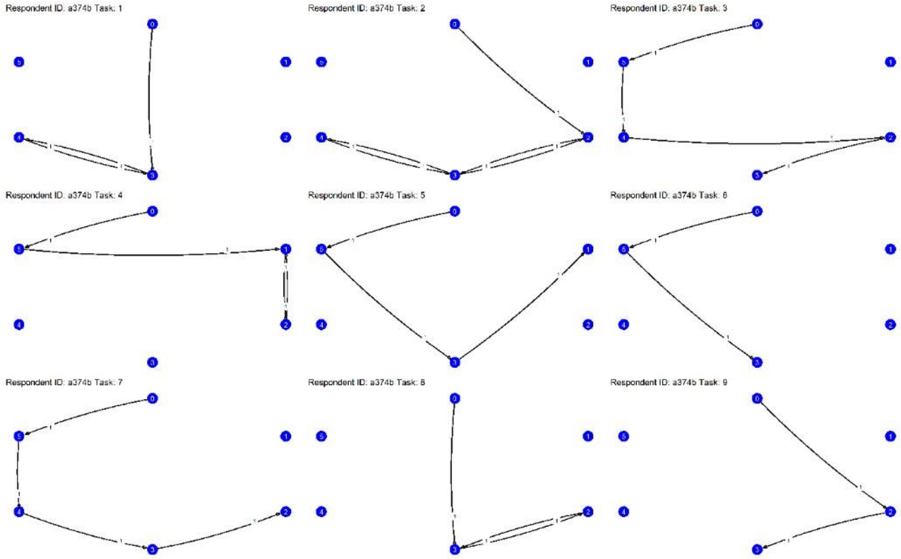
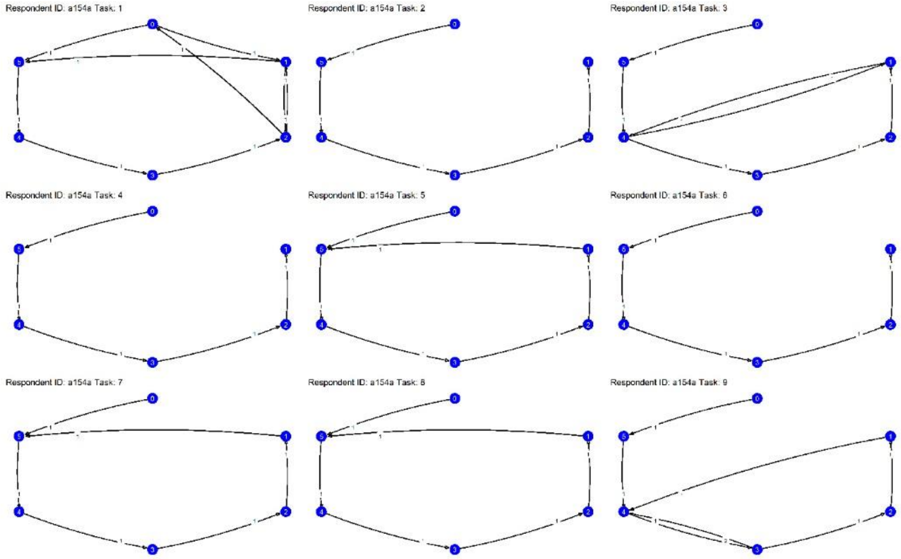
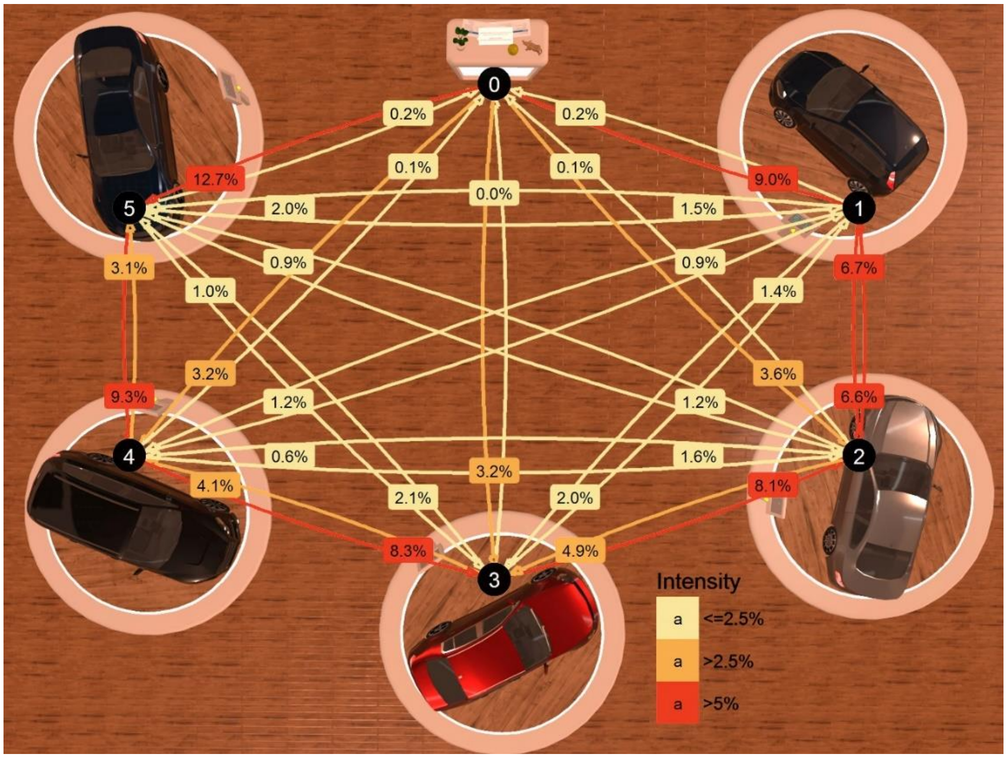

Respondent ID: a374b Task: 1
Respondent ID: a374b Task: 2
Respondent ID: a374b Task: 3
Respondent ID: a374b Task: 4
Respondent ID: a374b Task: 5
Respondent ID: a374b Task: 6
Respondent ID: a374b Task: 7
Respondent ID: a374b Task: 8
Respondent ID: a374b Task: 9

Figure 11: Path Graph per Respondent (Group 2)


Respondent ID: a154a Task: 1
Respondent ID: a154a Task: 2
Respondent ID: a154a Task: 3
Respondent ID: a154a Task: 4
Respondent ID: a154a Task: 5
Respondent ID: a154a Task: 6
Respondent ID: a154a Task: 7
Respondent ID: a154a Task: 8
Respondent ID: a154a Task: 9

By looking at the overall respondents' movement pattern, we identified that in 73.2% of the cases the respondents used a sequential pattern (mainly from right to left) rather than randomly moving from one car to another, as seen in Figure 10.

Figure 12: Aggregate Path Map

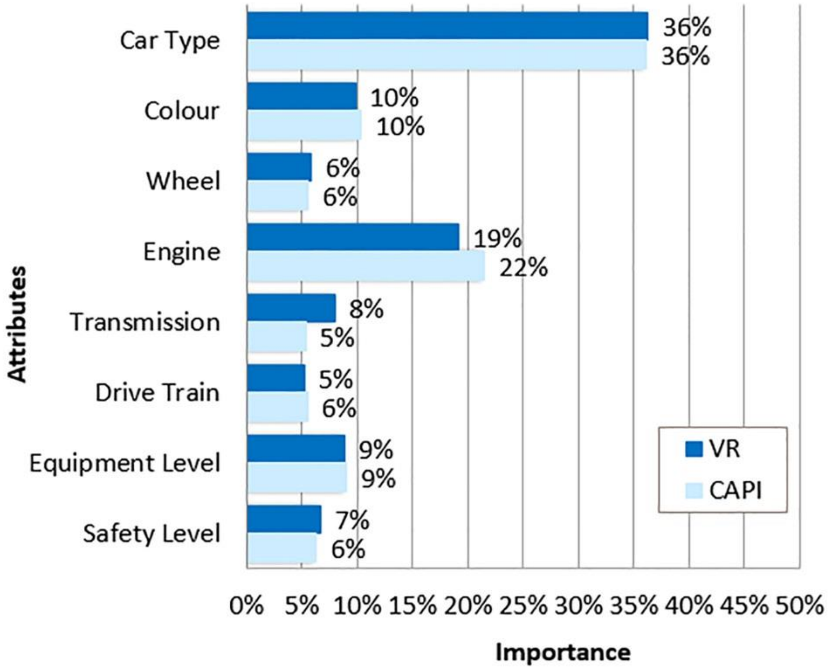


Conjoint Results

Attribute Importance

As summarized in Figure 12, the method (CAPI vs. VR) didn’t influence the decision drivers. The observed patterns are identical. All attributes such as car type, engine type, color, etc. are equally important in VR and CAPI.

Figure 13: Attribute Importance



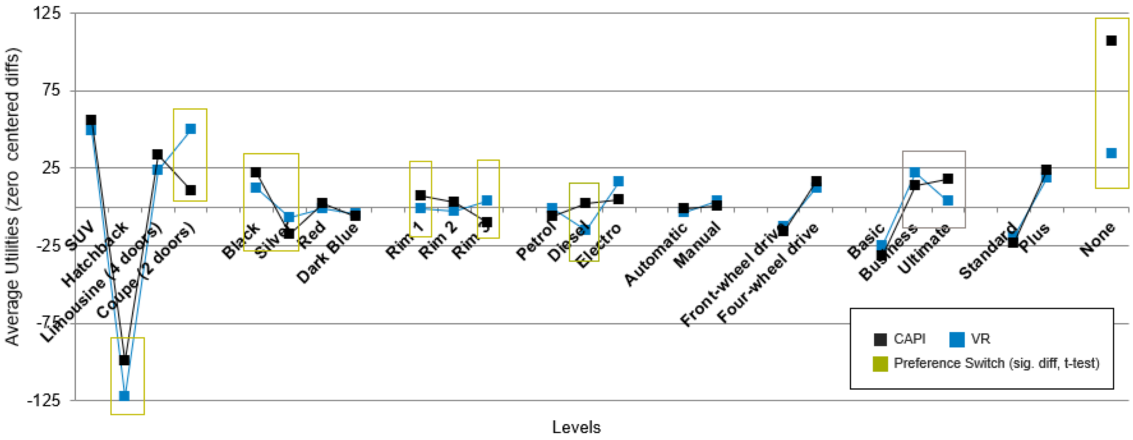
In both experiments, the respondent's decision was triggered by both visual and text-based attributes. This is especially important for VR, as the design of the experiment might have encouraged respondents to overweigh the visual attributes in their decision-making process, but this obviously did not happen.

Respondents' Preference

The respondents switched their preference for different specific car models between the experiments (VR vs. CAPI), as seen in Figure 13. Perhaps the aesthetic appeal of a coupé versus the more standard car types is more strongly portrayed in VR. This may also explain the statistical significant preference switch from wheel type 1 (in CAPI) to the sportier wheel type 3 (in VR).

In the case of diesel vs. electric, they were both equally chosen in CAPI. Yet, in VR the electric car was significantly preferred. It might be that the strong attraction of the coupé drove up the choice of electric in VR as these features were seen by participants as a compatible pair.

Figure 14: Scaled Conjoint Part-Worths (Zero-Centered Diffs)

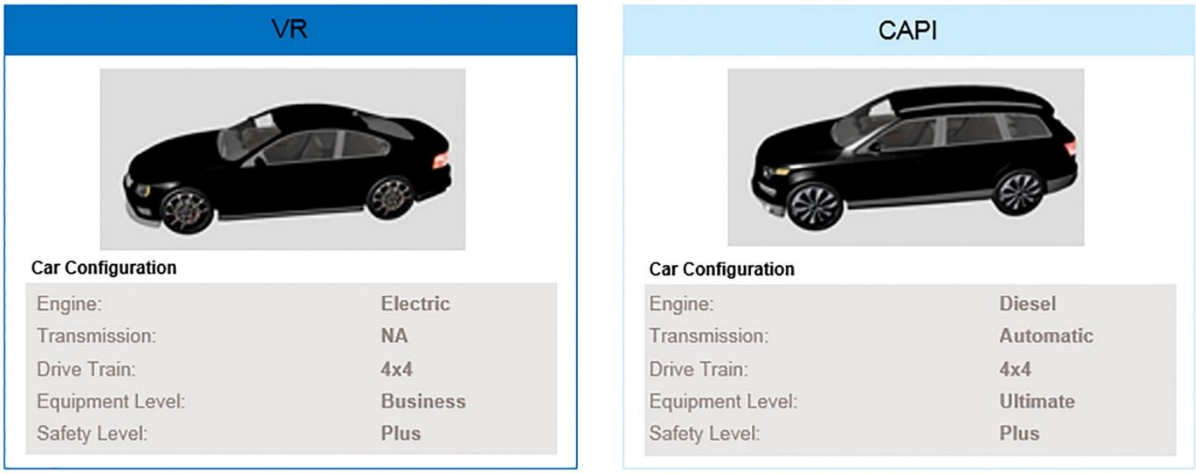


Interestingly, the hierarchy of choice in CAPI better reflects the actual purchasing trends on the market (e.g., Limousine preferred to Coupé, Source: JATO Dynamics Report, 2016) Considering also the fact that the time spent in VR for the decision-making process was lower than in CAPI, one might think that CAPI is more suited to instigate rational and conservative choices while VR encourages personal and emotional preferences to overrule conventions.

Ideal Car

We used an Excel-based simulator (share of preference) to search for the ideal car (the most preferred car based on individual preferences) in both of the two experiments. Below, the ideal car in VR and CAPI:

Figure 15: Ideal Car



VR

Car Configuration

Engine:	Electric
Transmission:	NA
Drive Train:	4x4
Equipment Level:	Business
Safety Level:	Plus

CAPI

Car Configuration

Engine:	Diesel
Transmission:	Automatic
Drive Train:	4x4
Equipment Level:	Ultimate
Safety Level:	Plus

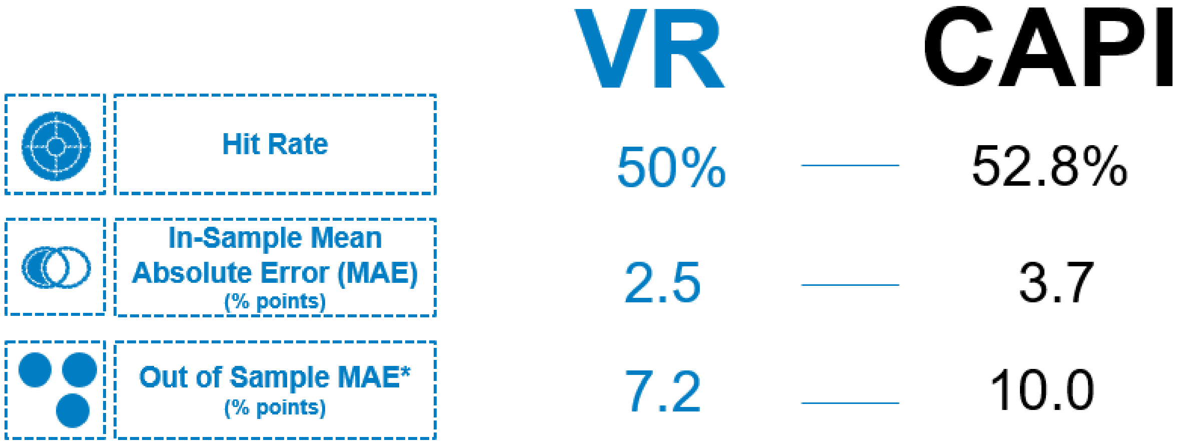
Using two different data collection techniques leads to two different results. When using CAPI, respondents' ideal car was a Diesel SUV while when using VR respondents preferred an Electric Coupé. It underlines once again that CAPI encourages respondents to rationally choose a car that would meet a lot of the practical demands of their life as the SUV does. On the other hand, the VR choice (the electric coupe) is a more aspirational one; perhaps VR allows participants to express their emotional preference, beyond their rational and day-by-day life demands.

Diagnostic

The in-sample and out-of-sample mean absolute errors are better for VR compared to CAPI. The in-sample MAE is based on 2 VR tasks, respectively 2 CAPI tasks. The out-of-sample MAE is based on 4 VR tasks, respectively 4 CAPI tasks. The reported out-of-sample MAE is calculated within the method (i.e., the VR model was used to predict the VR choices of the holdout sample). The average across all is reported in the Figure 15.

The hit rate is slightly better for CAPI compared to VR (52.8% vs. 50%).

Figure 16: Conjoint Diagnostic



	VR	CAPI
Hit Rate	50%	52.8%
In-Sample Mean Absolute Error (MAE) (% points)	2.5	3.7
Out of Sample MAE* (% points)	7.2	10.0

Conclusions and Managerial Recommendations

Logistically, it was no problem for respondents to properly use the VR equipment. Cognitively, VR felt like an intuitive test setting, with participants finding it easy and natural to go through the tasks. Hence, from the feasibility perspective, VR might be rather easily integrated as a market research technology.

The use of VR engaged the respondents more fully, creating a more satisfying survey environment than CAPI, which, in the end, translated into a better quality interview.

Therefore, this new technology appears to have the potential to aid recruitment and engage more difficult target groups, offering a promising outlook for recruiters and panel managers.

The immersion into a virtual auto showroom created for the respondents the perception of being physically present in a non-physical world. Perhaps, it is the sense of presence that activated the respondents' emotions. A VR-enhanced interview facilitates emotions to influence the respondents' judgments and choices and to express their preferences beyond conventions. In the end, all of this leads to a better data quality and a higher predictive power. Moreover, it was particularly interesting to notice that VR does not prompt the participants to oversimplify their choices by focusing only on the visual characteristics of a car. The other characteristics were equally important (same as seen in a CAPI environment).

This makes VR particularly interesting for design thinking/product optimization (co-creation platform) and concept testing when the success depends on a deep understanding of consumers' needs and preferences and where the traditional quantitative marketing research methods break (O'Hern and Rindfleisch, 2009; von Hippel, 2005).

It is important to underline that for this study, we used one of the most immersive VR headsets on the market, and we also created the conditions for an immersive experience (e.g., a quiet location, limited interactions between external factors and respondent during the VR interview). The results of this research might not be reproduced in less immersive test design.

This research concentrated on understanding the impact of VR on respondents' preference and behaviour, therefore further validation against revealed behaviour and buying decisions is needed to consider using VR for demand estimation studies or even forecasting. Besides, this study was focused on one category (automotive) and conducted in one country. Conducting similar research on different categories and in other countries would help understanding its scalability (applicability across multiple industries and regions).

Alexandra Chirilov

TOO MUCH INFORMATION?: THE CURIOUS CASE OF AUGMENTED MAXDIFF

JACKIE GUTHART
CURTIS FRAZIER
RAMAN SAINI
RADIUS GLOBAL MARKET RESEARCH

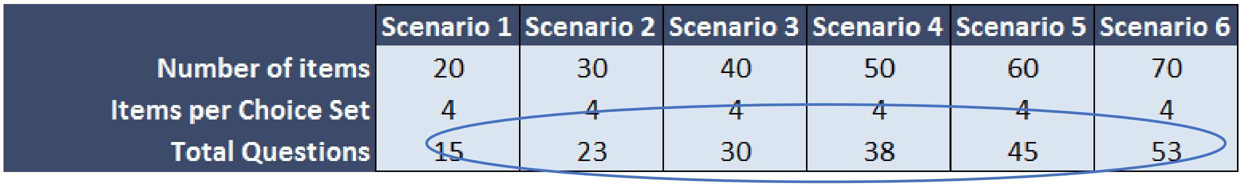
ABSTRACT

Augmented MaxDiff has become an effective and increasingly used variant for dealing with large numbers of attributes. While Augmented MaxDiff is efficient and effective at recovering individual-level preferences, there is a little talked about issue in the design setup and estimation of utilities. We will illustrate the issue of "too much information" and discuss design decisions to mitigate utilities estimation problems.

INTRODUCTION

When estimating MaxDiff data, to get good individual-level estimates, we require at least 3 exposures per item using hierarchical Bayesian estimation. Under this condition, as the number of items tested in the MaxDiff increase, the number of questions required per respondent also increases, which leads to increased respondent fatigue and survey length. Table 1 shows how many questions we would need to ask for different number of items for a 3 exposures per item setup.

Table 1: Questions Needed for Different Number of Items



	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
Number of items	20	30	40	50	60	70
Items per Choice Set	4	4	4	4	4	4
Total Questions	15	23	30	38	45	53

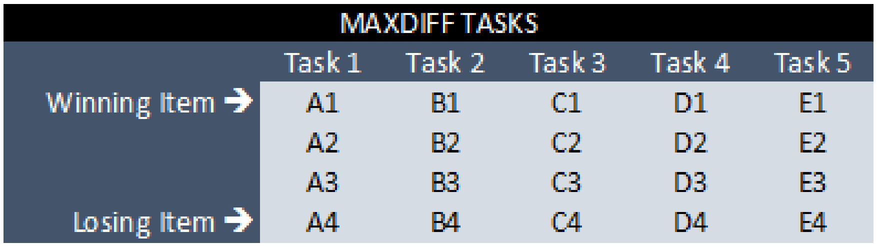
To counteract the issue, we could relax the number of exposures and ask a Sparse MaxDiff with 1 exposure per item, however, we would lose precision on individual-level estimates.

We typically would do a traditional MaxDiff when we have less than 30 items. Augmented MaxDiff becomes a preferred technique when we are testing approximately 30 to 60 items. When we have more than 60 items, techniques such as Express MaxDiff and Bandit MaxDiff are more appropriate.

In Augmented MaxDiff, we ask all the items to the respondents in a Sparse MaxDiff, exposing each item just one time. Then, like the name suggests, we augment the information we get from the Sparse MaxDiff with information we get from external questions. These external questions could be in the form of a Q-Sort exercise, an Adaptive MaxDiff, or ranking question(s) of the items selected as "best" and optionally the items selected as "worst." Our preferred method of augmentation is asking ranking questions. Ranking questions give us more information about the "best" items, which the client is typically most interested in. The ranking of "worst" items is optional and could be useful for doing a post hoc analysis where we require good individual-level estimates on "worst" items, for example a segmentation. Research has shown that augmenting on both top and bottom items is best for doing any follow-up analysis (Jones and Yeh, 2013). The decision about the amount of information to put into the model from the augmented questions is up to the researcher.

Table 2a and Table 2b below show an example Augmented MaxDiff task. Figure 2a shows an example Sparse MaxDiff design with 20 items (A1 through E4). It also shows which item a respondent selected as "best" and which was selected as "worst."

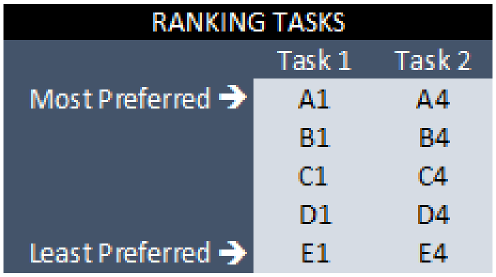
Table 2a: Sparse MaxDiff Task



MAXDIFF TASKS					
	Task 1	Task 2	Task 3	Task 4	Task 5
Winning Item ➔	A1	B1	C1	D1	E1
	A2	B2	C2	D2	E2
	A3	B3	C3	D3	E3
Losing Item ➔	A4	B4	C4	D4	E4

Table 2b illustrates the two follow-up ranking tasks that the respondent must complete. In Task 1, respondents are shown the 5 items they selected as "best" in the original MaxDiff tasks. And, in Task 2, respondents are asked to rank the 5 items they selected as "worst" in the original MaxDiff tasks. It also shows the rank ordering of items given by the respondent in each ranking task.

Table 2b: Ranking Tasks

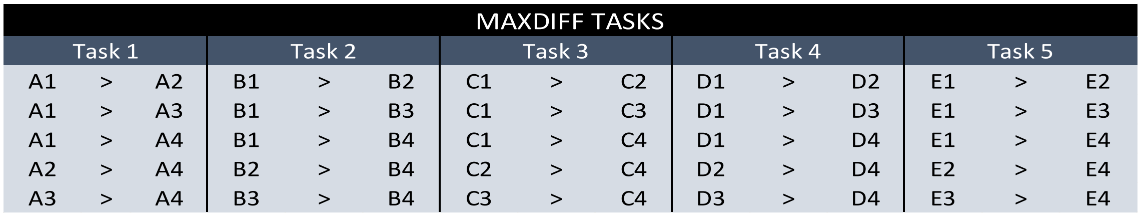


RANKING TASKS		
	Task 1	Task 2
Most Preferred ➔	A1	A4
	B1	B4
	C1	C4
	D1	D4
Least Preferred ➔	E1	E4

Analysis Considerations

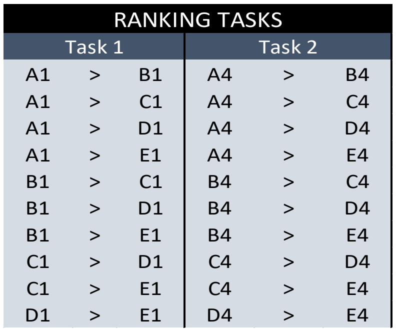
There are two key decisions the researcher must make when specifying the model. The first decision is the amount of information that goes into the model. From the example in Table 2a and Table 2b, we have 25 pairs of known preferences from the Sparse MaxDiff (shown in Table 3a) before we consider the ranking tasks, which could add another 20 pairs from the ranking task (Table 3b). We've not included any inferred comparisons, for example A1 wins against B1 in the ranking tasks implying A1 wins against all items that B1 is winning against in MaxDiff Task 2.

Table 3a: Paired Comparisons from the MaxDiff Tasks



MAXDIFF TASKS														
Task 1			Task 2			Task 3			Task 4			Task 5		
A1	>	A2	B1	>	B2	C1	>	C2	D1	>	D2	E1	>	E2
A1	>	A3	B1	>	B3	C1	>	C3	D1	>	D3	E1	>	E3
A1	>	A4	B1	>	B4	C1	>	C4	D1	>	D4	E1	>	E4
A2	>	A4	B2	>	B4	C2	>	C4	D2	>	D4	E2	>	E4
A3	>	A4	B3	>	B4	C3	>	C4	D3	>	D4	E3	>	E4

Table 3b: Paired Comparisons from the Ranking Tasks



RANKING TASKS					
Task 1			Task 2		
A1	>	B1	A4	>	B4
A1	>	C1	A4	>	C4
A1	>	D1	A4	>	D4
A1	>	E1	A4	>	E4
B1	>	C1	B4	>	C4
B1	>	D1	B4	>	D4
B1	>	E1	B4	>	E4
C1	>	D1	C4	>	D4
C1	>	E1	C4	>	E4
D1	>	E1	D4	>	E4

The second decision concerns the number of total iterations that we let the estimation run for (initial iterations so that it reaches convergence, and iterations that we use for draws for averaging the estimates).

Based on numerous Augmented MaxDiff analysis runs, we've seen that these two critical considerations play a big role in changing our utility estimates and potentially impacting business decisions. There appears to be a balancing act between the amount and specification of information that is provided in the input file and the number of iterations that the model runs for.

Introduction to the Data

For our analyses we looked at datasets from 3 past studies that we conducted. These three studies were chosen out of dozens of Augmented MaxDiff projects because they represent different numbers of items, design sizes, augmentation specifications, and industries. The specifications of the 3 studies were as shown in Table 4.

Table 4: Data Specifications for Case Studies

	Drug Store Shopping Landscape Concepts Test	Children Pain Relief Claims Test	Laptop Features Claims Test
# Items	29	44	55
# Items per screen	2 or 3	4	2 or 3
# Screens	10	11	19
Ranking questions	10 Best & 10 Worst	11 Best & 11 Worst	12 Best & 12 Worst
# Respondents	500	500	500

Analytics Plan

We estimated models in which each of our critical decisions varied. We varied the level of pairs to include in the estimation of the model, and we also varied the number of iterations. For each of the different models that we ran, we evaluated the:

- Aggregate level rescaled MaxDiff scores,
- Average RLH,
- Distribution of individual-level preferences focusing on proportion of items that were considered having extreme scores (Rescaled MaxDiff utility of <10 or >90), and
- Differences in expected vs. modeled ranking of the best and worst ranked items.

Augmentation Schemes

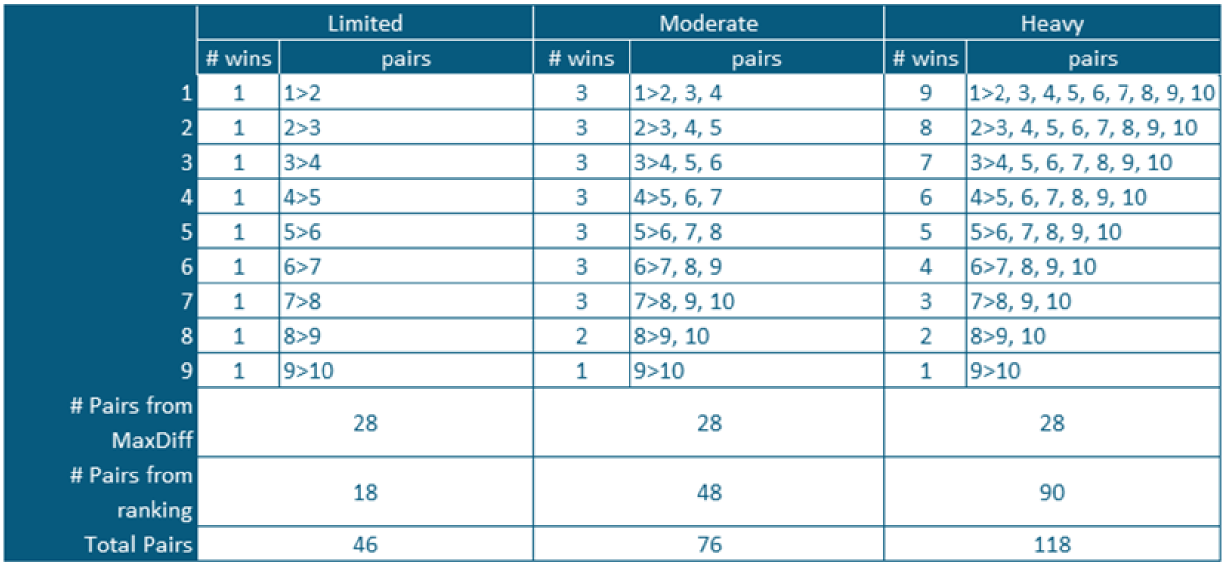
We estimated models with 4 levels of augmentation:

- Sparse, or no augmentation
- Limited
- Moderate
- Heavy

An example of the pairs of ranked items that were compared in each level of augmentation for the drug store study, which had 29 items with rankings of the 10 items selected best and 10 items selected worst, is illustrated in Table 5.

Note: Table 5 just shows the examples of the "best" rankings—the total number of pairs is double this because we also have "worst" rankings.

Table 5: Different Levels of Augmentation—Drug Store



	Limited		Moderate		Heavy	
	# wins	pairs	# wins	pairs	# wins	pairs
1	1	1>2	3	1>2, 3, 4	9	1>2, 3, 4, 5, 6, 7, 8, 9, 10
2	1	2>3	3	2>3, 4, 5	8	2>3, 4, 5, 6, 7, 8, 9, 10
3	1	3>4	3	3>4, 5, 6	7	3>4, 5, 6, 7, 8, 9, 10
4	1	4>5	3	4>5, 6, 7	6	4>5, 6, 7, 8, 9, 10
5	1	5>6	3	5>6, 7, 8	5	5>6, 7, 8, 9, 10
6	1	6>7	3	6>7, 8, 9	4	6>7, 8, 9, 10
7	1	7>8	3	7>8, 9, 10	3	7>8, 9, 10
8	1	8>9	2	8>9, 10	2	8>9, 10
9	1	9>10	1	9>10	1	9>10
# Pairs from MaxDiff	28		28		28	
# Pairs from ranking	18		48		90	
Total Pairs	46		76		118	

The sparse models only include the 28 pairs from the MaxDiff questions. In limited augmentation we have ranking items compared with the adjacent ranked items—item 1 is better than item 2, item 2 is better than item 3, and so on. Moderate augmentation has each item compared with the next 3 items ranked—item 1 is better than 2, 3, and 4; etc. Finally, in heavy augmentation all items are compared to all items they're winning against in the augmentation questions.

Aggregate-Level Estimation

We estimated the aggregate-level utilities of the different level of augmentation for each of the 3 studies using the Sawtooth default number of iterations (20,000 initial and 10,000 draws) using HB estimation. The results of the estimation of each of the 3 studies are shown in Figures 1, 2, and 3.

Figure 1: Aggregate-Level HB Utilities Default Iterations—Drug Store

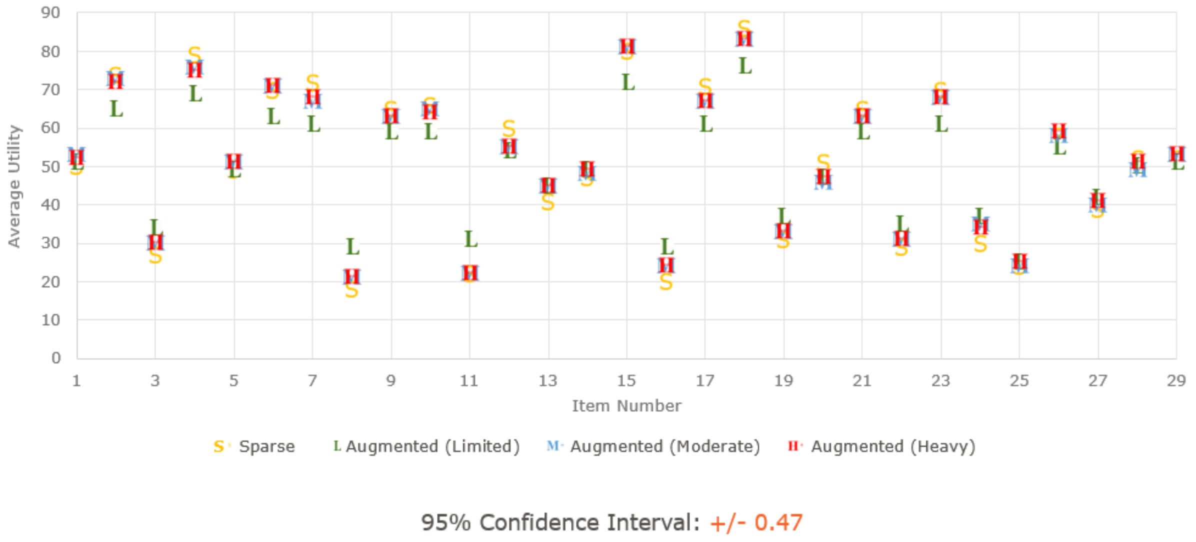


Figure 2: Aggregate-Level HB Utilities Default Iterations—Pain Reliever

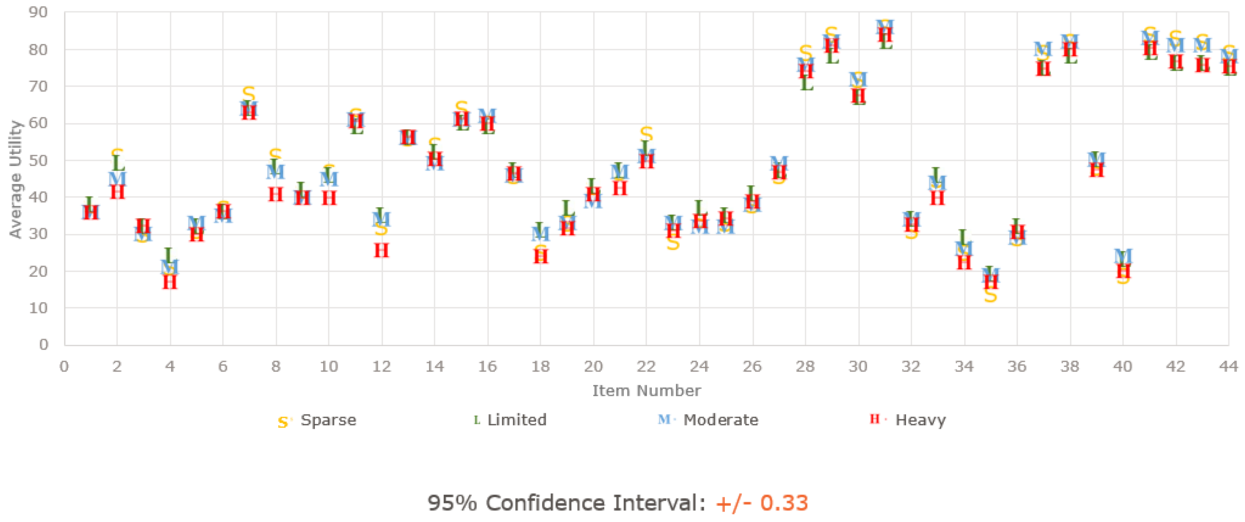
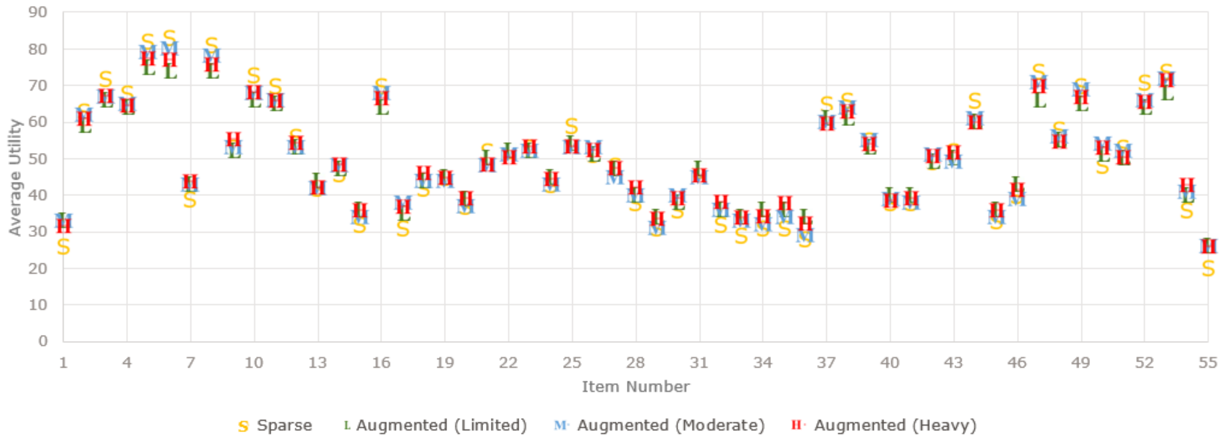


Figure 3: Aggregate-Level HB Utilities Default Iterations—Laptops



95% Confidence Interval: +/- 0.22

At the aggregate level, the results are exceptionally consistent—the winning items are winning, and the losing items are losing. Regardless of approach, the recommendations to the client do not change. With a 95% confidence interval, the average estimated utilities for the 3 studies are within +/- 0.47 across all studies, meaning the aggregate-level MaxDiff estimates are very similar irrespective of level of augmentation.

We also looked at the average RLH for each of these runs—shown below in Table 6.

Table 6: Default Iterations Average RLH

	Sparse	Limited	Moderate	Heavy
Drug Store	0.75	0.57	0.79	0.91
Pain Reliever	0.83	0.65	0.90	0.92
Laptops	0.74	0.59	0.61	0.87

The more information we augment the model with, the higher the average RLH. If higher average RLH were to be taken as an indicator of better model fit, we'd be inclined to always specify models with heavy augmentation. However, at some point a higher RLH isn't an indicator of a better model, and average RLH scores of 0.92 aren't a realistic expectation of human responses. It's an indicator of overfitting.

Investigation of the Underlying Issue

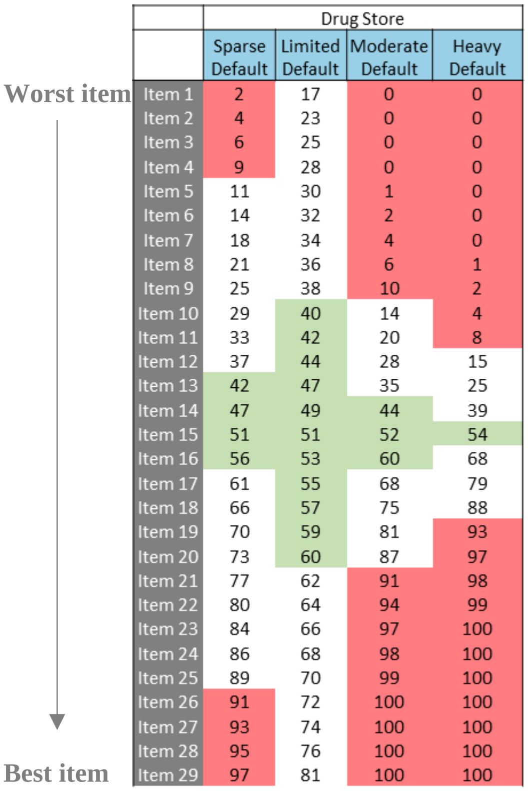
When we take a closer look at individual-level utilities, we see that there is a pretty drastic skew particularly in cases where there is more information going into the model.

Table 7 shows average utilities after sorting each item from worst to best (item 1 is each person's worst item). When we look at average utilities it seems to balance out since we typically have heterogeneity in the data, but Table 7, below, indicates that at the individual level, our predictions might not be doing a great job. Heavier augmentation yields a more drastic skew.

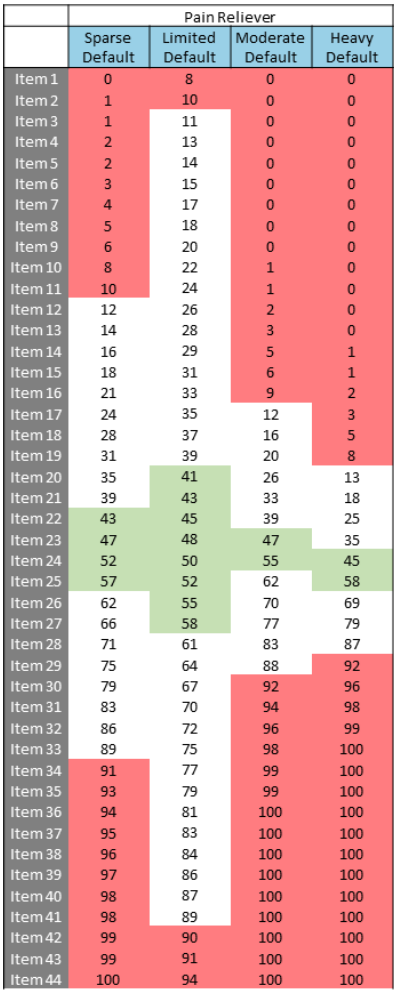
Inspecting the individual-level data, we noticed a strong skew towards more extreme scores as we increased the level of augmentation. Table 7 illustrates this skew. In these tables, we have highlighted any scores in the top and bottom deciles (>=90 or <=10). We would expect that these two deciles would account for about 20% of items; with high levels of augmentation we see that these two deciles can account for over 80% of all items.

Table 7: Average Rescaled MaxDiff Utilities by Augmentation Level

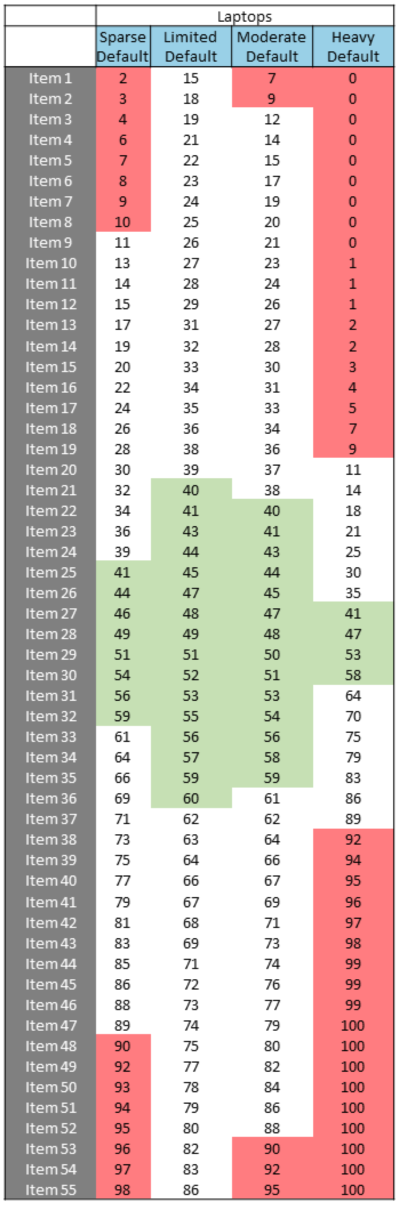
Red: 0–10; 90–100
Green: 40–60


Worst item
Best item

	Drug Store			
	Sparse Default	Limited Default	Moderate Default	Heavy Default
Item 1	2	17	0	0
Item 2	4	23	0	0
Item 3	6	25	0	0
Item 4	9	28	0	0
Item 5	11	30	1	0
Item 6	14	32	2	0
Item 7	18	34	4	0
Item 8	21	36	6	1
Item 9	25	38	10	2
Item 10	29	40	14	4
Item 11	33	42	20	8
Item 12	37	44	28	15
Item 13	42	47	35	25
Item 14	47	49	44	39
Item 15	51	51	52	54
Item 16	56	53	60	68
Item 17	61	55	68	79
Item 18	66	57	75	88
Item 19	70	59	81	93
Item 20	73	60	87	97
Item 21	77	62	91	98
Item 22	80	64	94	99
Item 23	84	66	97	100
Item 24	86	68	98	100
Item 25	89	70	99	100
Item 26	91	72	100	100
Item 27	93	74	100	100
Item 28	95	76	100	100
Item 29	97	81	100	100



	Pain Reliever			
	Sparse Default	Limited Default	Moderate Default	Heavy Default
Item 1	0	8	0	0
Item 2	1	10	0	0
Item 3	1	11	0	0
Item 4	2	13	0	0
Item 5	2	14	0	0
Item 6	3	15	0	0
Item 7	4	17	0	0
Item 8	5	18	0	0
Item 9	6	20	0	0
Item 10	8	22	1	0
Item 11	10	24	1	0
Item 12	12	26	2	0
Item 13	14	28	3	0
Item 14	16	29	5	1
Item 15	18	31	6	1
Item 16	21	33	9	2
Item 17	24	35	12	3
Item 18	28	37	16	5
Item 19	31	39	20	8
Item 20	35	41	26	13
Item 21	39	43	33	18
Item 22	43	45	39	25
Item 23	47	48	47	35
Item 24	52	50	55	45
Item 25	57	52	62	58
Item 26	62	55	70	69
Item 27	66	58	77	79
Item 28	71	61	83	87
Item 29	75	64	88	92
Item 30	79	67	92	96
Item 31	83	70	94	98
Item 32	86	72	96	99
Item 33	89	75	98	100
Item 34	91	77	99	100
Item 35	93	79	99	100
Item 36	94	81	100	100
Item 37	95	83	100	100
Item 38	96	84	100	100
Item 39	97	86	100	100
Item 40	98	87	100	100
Item 41	98	89	100	100
Item 42	99	90	100	100
Item 43	99	91	100	100
Item 44	100	94	100	100



	Laptops			
	Sparse Default	Limited Default	Moderate Default	Heavy Default
Item 1	2	15	7	0
Item 2	3	18	9	0
Item 3	4	19	12	0
Item 4	6	21	14	0
Item 5	7	22	15	0
Item 6	8	23	17	0
Item 7	9	24	19	0
Item 8	10	25	20	0
Item 9	11	26	21	0
Item 10	13	27	23	1
Item 11	14	28	24	1
Item 12	15	29	26	1
Item 13	17	31	27	2
Item 14	19	32	28	2
Item 15	20	33	30	3
Item 16	22	34	31	4
Item 17	24	35	33	5
Item 18	26	36	34	7
Item 19	28	38	36	9
Item 20	30	39	37	11
Item 21	32	40	38	14
Item 22	34	41	40	18
Item 23	36	43	41	21
Item 24	39	44	43	25
Item 25	41	45	44	30
Item 26	44	47	45	35
Item 27	46	48	47	41
Item 28	49	49	48	47
Item 29	51	51	50	53
Item 30	54	52	51	58
Item 31	56	53	53	64
Item 32	59	55	54	70
Item 33	61	56	56	75
Item 34	64	57	58	79
Item 35	66	59	59	83
Item 36	69	60	61	86
Item 37	71	62	62	89
Item 38	73	63	64	92
Item 39	75	64	66	94
Item 40	77	66	67	95
Item 41	79	67	69	96
Item 42	81	68	71	97
Item 43	83	69	73	98
Item 44	85	71	74	99
Item 45	86	72	76	99
Item 46	88	73	77	99
Item 47	89	74	79	100
Item 48	90	75	80	100
Item 49	92	77	82	100
Item 50	93	78	84	100
Item 51	94	79	86	100
Item 52	95	80	88	100
Item 53	96	82	90	100
Item 54	97	83	92	100
Item 55	98	86	95	100

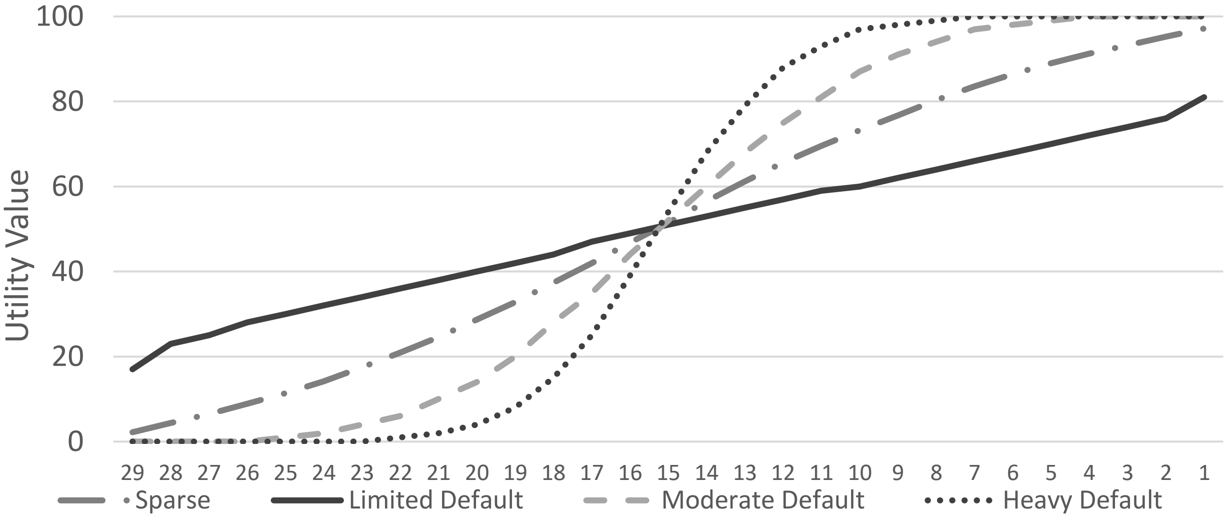
Note, the utilities are scaled using the probability-based rescaling procedure (The MaxDiff System Technical Paper, p.13). To convert the raw utilities to 0–100 scale the following transformation was performed:

$$\textbf{Rescaled Utility for item i=} \frac{\boldsymbol{e^{Ui}}}{(\boldsymbol{e^{Ui}+a-1})}$$

Ui = zero-centered raw logit weight for item I, e^{Ui} is equivalent to taking the antilog of Ui, a = Number of items shown per set.

Figure 4 shows the same data in a more visual way to better understand what the distribution looks like. It is clear from this graph that increased information in the model results in an increased number of "extreme" scores. After looking at distribution from multiple standard MaxDiff studies we found that on average there are ~20% of items that fall in the "extreme" range.

Figure 4: A Graphical View of the Distribution—Drug Store



Predictive Validity

To better understand predictive validity of the models at the aggregate level, we first investigated the median aggregate rankings of the items respondents selected as best in the Sparse MaxDiff. Table 8 shows the median rank of where the estimated utility lies for the item each respondent ranked as 1st best, 2nd best, and so on, in the follow-up ranking task of all the items the respondent selected as best in the Sparse MaxDiff.

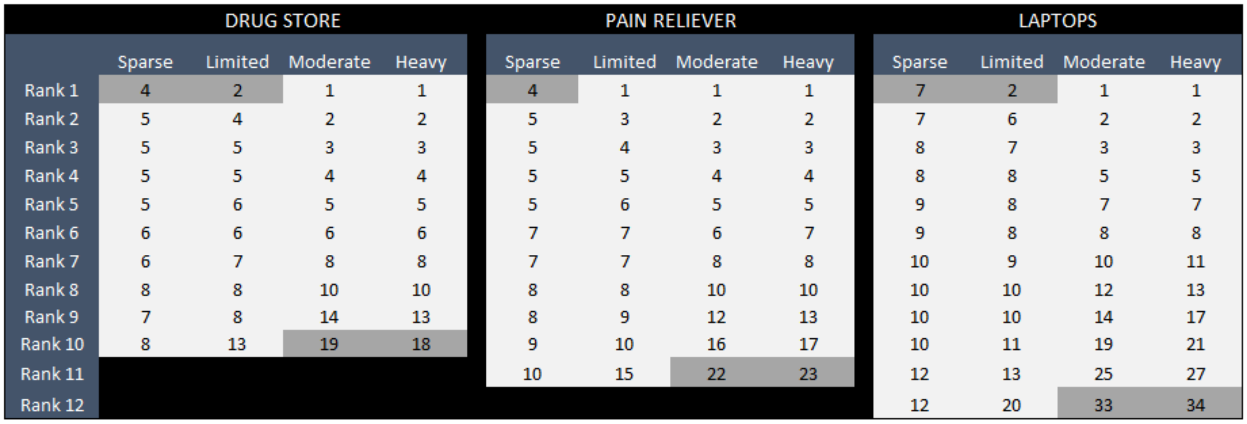
There are some things we know based off how a respondent ranks the items:

1. The item ranked as best in the ranking question is the respondent's top item (we also know the absolute ranking for the worst item).
2. In the case of the drug store scenario, for example, where a respondent saw 2–3 items per screen with 10 screens and a ranking of 10 items, we know the 2nd best item from the ranking is truly the respondents 2nd best, 3rd best, or 4th best. If for example someone had their 3 favorite things in the same MaxDiff task, then it's possible their 2nd best ranked item is only their 4th best item.

3. While we don’t know the exact ranking position of the 10th best ranked item, we know there should be variation in where it falls between rank position 10 and rank position 27, with the majority falling between rank position 10 and rank position 20.
4. We know the preference order of all the items ranked best and similarly all the items ranked worst.

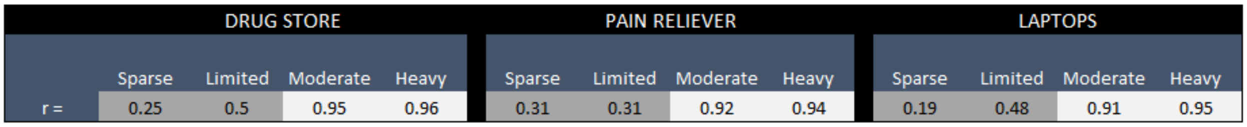
In Table 9, we’ve highlighted those predicted rankings that violate what we either know or strongly suspect the true rankings should be. For example, Sparse MaxDiff is consistently predicting what we know to be the best of the best, as only the 4th most preferred item. Moderate and Heavy levels of augmentation are consistently predicting what we *believe* to be, at least, a slightly preferred item to be worse than average. Table 9 shows average correlations at the individual level of the rankings of the items selected as best, to the known order that best ranked 1 is better than best ranked 2, and so on.

Table 8: Median Rankings of Items in the Ranking Test of Items Selected as Best



	DRUG STORE				PAIN RELIEVER				LAPTOPS			
	Sparse	Limited	Moderate	Heavy	Sparse	Limited	Moderate	Heavy	Sparse	Limited	Moderate	Heavy
Rank 1	4	2	1	1	4	1	1	1	7	2	1	1
Rank 2	5	4	2	2	5	3	2	2	7	6	2	2
Rank 3	5	5	3	3	5	4	3	3	8	7	3	3
Rank 4	5	5	4	4	5	5	4	4	8	8	5	5
Rank 5	5	6	5	5	5	6	5	5	9	8	7	7
Rank 6	6	6	6	6	7	7	6	7	9	8	8	8
Rank 7	6	7	8	8	7	7	8	8	10	9	10	11
Rank 8	8	8	10	10	8	8	10	10	10	10	12	13
Rank 9	7	8	14	13	8	9	12	13	10	10	14	17
Rank 10	8	13	19	18	9	10	16	17	10	11	19	21
Rank 11					10	15	22	23	12	13	25	27
Rank 12									12	20	33	34

Table 9: Average Individual Level Correlation Between Item Rankings
and Predicted Rankings



	DRUG STORE				PAIN RELIEVER				LAPTOPS			
	Sparse	Limited	Moderate	Heavy	Sparse	Limited	Moderate	Heavy	Sparse	Limited	Moderate	Heavy
r =	0.25	0.5	0.95	0.96	0.31	0.31	0.92	0.94	0.19	0.48	0.91	0.95

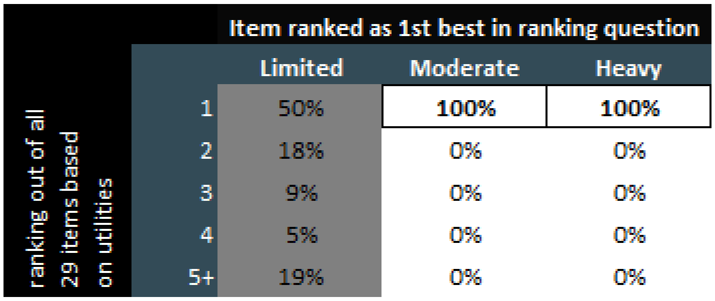
The Median aggregate rankings in Table 8 suggest that items are being modeled in the correct rank order. However, looking at Table 9 we see that this finding doesn’t hold true at the individual level—especially in the case of limited augmentation. Heavier augmentation shows rankings are highly correlated at the individual level, but it yields unrealistically poor rankings (seen in Table 8) of the 9th+ best items.

Looking at the Median aggregate ranking only tells part of the story. If we take the drug store example and further look at the distribution of the predicted rank position of the 1st best item in the rank question, 2nd best item in the rank question, and 10th best item in the rank question, the problem at the individual level further emerges. We know the 1st best ranked item is the respondent’s best item overall so in theory 100% of people should have the highest predicted MaxDiff utility for this item. This is true in the case of Moderate and Heavy augmentation, but for Limited, only 50% of respondents’ 1st best item was accurately being predicted as best as shown in Table 10 even though the median ranking in Table 8 is 1.

We know the 2nd best item from the ranking question should always be the 2nd–4th best item overall. Looking at Table 11 starts to uncover issues with Moderate and Heavy augmentation which strongly suggest these models are overfitting to the ranking questions. We see that 100% of respondents' 2nd best item from the ranking is being predicted as the 2nd best item overall.

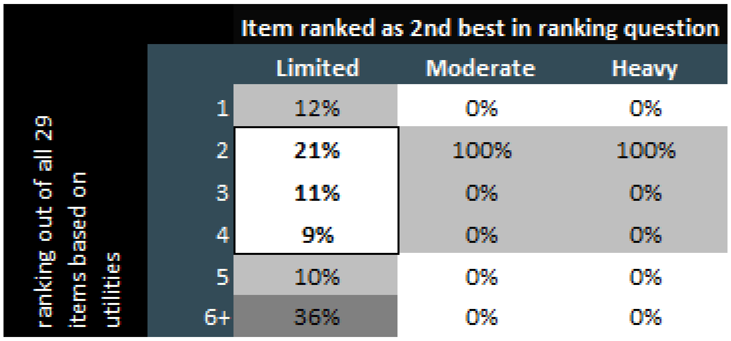
Looking at the distribution of the 10th best item in Table 12 it's apparent that Limited is doing a poor job since it predicts the 10th best item from the ranking to win over items we know it is worse than for 25% of respondents. On the opposite side of the spectrum, Moderate and Heavy augmentation are predicting the 10th best ranked item to be worse than the losing items in the MaxDiff for over 30% of respondents—this is likely too high.

Table 10: Distribution of Ranking Based on Utility of the 1st Best Ranked Item



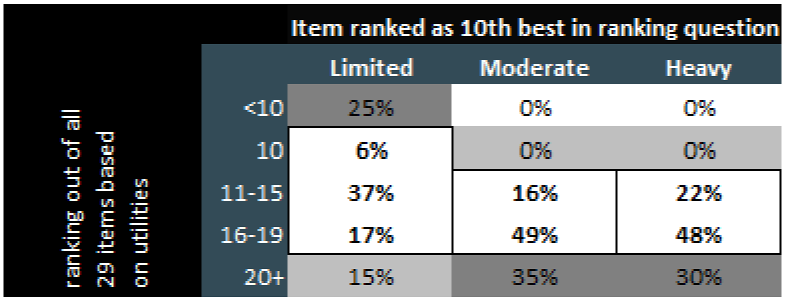
		Item ranked as 1st best in ranking question		
		Limited	Moderate	Heavy
ranking out of all 29 items based on utilities	1	50%	100%	100%
	2	18%	0%	0%
	3	9%	0%	0%
	4	5%	0%	0%
	5+	19%	0%	0%

Table 11: Distribution of Ranking Based on Utility of the 2nd Best Ranked Item



		Item ranked as 2nd best in ranking question		
		Limited	Moderate	Heavy
ranking out of all 29 items based on utilities	1	12%	0%	0%
	2	21%	100%	100%
	3	11%	0%	0%
	4	9%	0%	0%
	5	10%	0%	0%
	6+	36%	0%	0%

Table 12: Distribution of Ranking Based on Utility of the 10th Best Ranked Item



		Item ranked as 10th best in ranking question		
		Limited	Moderate	Heavy
ranking out of all 29 items based on utilities	<10	25%	0%	0%
	10	6%	0%	0%
	11-15	37%	16%	22%
	16-19	17%	49%	48%
	20+	15%	35%	30%

Impact of Iterations

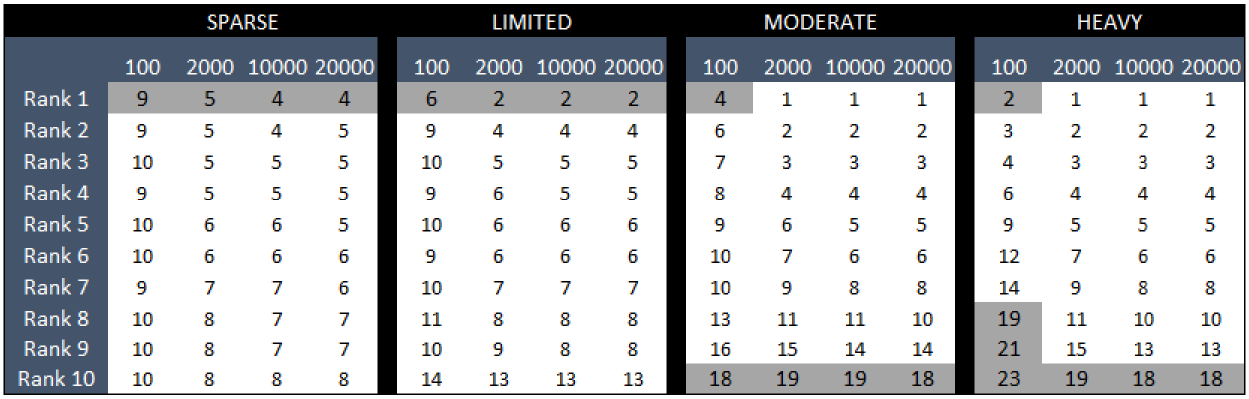
The other aspect a modeler has control over is the number of iterations to include when running a MaxDiff. We continue with the drug store example which shows that adding iterations can result in even more "extreme" scores. In Table 13, below, Limited isn't getting any extreme scores predicted even after the model converges. Moderate appears to have a nice distribution early on, somewhere between 2,000–10,000 initial iterations; however, the model still hasn't converged at this point—once it approaches convergence the number of extreme scores begins to get unwieldy. A similar story happens with Heavy augmentation but even more extreme. We hypothesize this happens because we have blown out the ranking task so much that we are introducing so many more wins (from best ranking) and losses (from worst ranking) that it overpowers the MaxDiff tasks and takes over the model. We observed the same pattern in the Pain Reliever and Laptop models as well (not shown here).

Table 13: Average Rescaled MaxDiff Utilities by Augmentation Level—Drug Store

		Sparse				Augmented- Limited				Augmented- Moderate				Augmented- Heavy			
	Initial	100	2000	10000	20000	100	2000	10000	20000	100	2000	10000	20000	100	2000	10000	20000
	Draws	50	1000	5000	10000	50	1000	5000	10000	50	1000	5000	10000	50	1000	5000	10000
Worst item	Item 1	31	3	2	2	32	16	17	17	31	4	0	0	17	0	0	0
	Item 2	34	6	4	4	35	22	22	23	34	8	1	0	23	0	0	0
	Item 3	37	9	6	6	37	24	25	25	37	12	2	0	27	0	0	0
	Item 4	39	12	8	9	39	27	28	28	39	15	3	0	30	1	0	0
	Item 5	40	15	11	11	41	29	30	30	40	18	6	1	33	2	0	0
	Item 6	41	18	14	14	42	31	32	32	41	22	9	2	35	3	0	0
	Item 7	43	21	17	18	43	34	34	34	43	25	12	4	37	5	0	0
	Item 8	44	24	21	21	44	36	36	36	44	29	16	6	39	8	1	1
	Item 9	45	28	24	25	45	38	38	38	45	32	21	10	41	12	3	2
	Item 10	46	32	29	29	46	40	40	40	46	35	26	14	42	17	5	4
	Item 11	46	35	33	33	47	42	42	42	47	38	31	20	44	23	10	8
	Item 12	47	39	37	37	48	44	44	44	48	42	35	28	45	29	17	15
	Item 13	48	44	42	42	48	47	46	47	48	45	41	35	47	36	27	25
	Item 14	49	48	47	47	49	49	49	49	49	48	46	44	49	43	39	39
	Item 15	50	52	52	51	50	51	51	51	50	51	51	52	50	51	53	54
	Item 16	51	56	56	56	51	53	53	53	51	54	56	60	52	58	66	68
	Item 17	52	60	61	61	52	55	55	55	52	57	62	68	53	66	77	79
	Item 18	52	63	66	66	52	57	57	57	53	60	67	75	55	73	86	88
	Item 19	53	67	70	70	53	59	59	59	53	63	71	81	56	79	92	93
	Item 20	54	71	74	73	54	61	60	60	54	66	76	87	58	84	95	97
	Item 21	55	74	77	77	55	63	62	62	55	70	80	91	60	89	98	98
	Item 22	56	77	80	80	56	65	64	64	56	73	85	94	61	92	99	99
	Item 23	57	80	83	84	57	67	66	66	57	76	88	97	63	95	100	100
	Item 24	59	83	87	86	58	69	68	68	59	79	91	98	65	97	100	100
	Item 25	60	86	89	89	60	71	70	70	60	81	94	99	68	98	100	100
	Item 26	62	88	92	91	61	73	72	72	61	85	96	100	70	99	100	100
	Item 27	63	91	94	93	63	75	75	74	63	88	98	100	73	100	100	100
	Item 28	66	93	96	95	65	78	77	76	66	92	99	100	77	100	100	100
Best item	Item 29	70	96	97	97	68	81	81	81	69	95	100	100	82	100	100	100

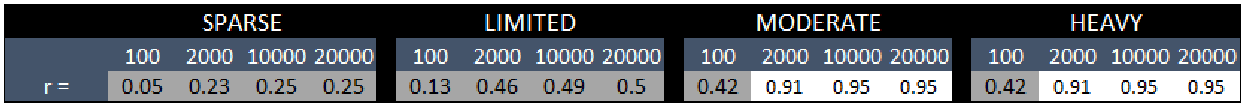
Similarly, we can look at rank order prediction by number of iterations. Looking at Tables 14 and 15, below, we see that fewer iterations leads to more inaccuracy of predicting the ranked items in the correct order, especially at the individual level.

Table 14: Median Rankings of Items Selected as Best in the Ranking Question by Iterations—Drug Store



	SPARSE				LIMITED				MODERATE				HEAVY			
	100	2000	10000	20000	100	2000	10000	20000	100	2000	10000	20000	100	2000	10000	20000
Rank 1	9	5	4	4	6	2	2	2	4	1	1	1	2	1	1	1
Rank 2	9	5	4	5	9	4	4	4	6	2	2	2	3	2	2	2
Rank 3	10	5	5	5	10	5	5	5	7	3	3	3	4	3	3	3
Rank 4	9	5	5	5	9	6	5	5	8	4	4	4	6	4	4	4
Rank 5	10	6	6	5	10	6	6	6	9	6	5	5	9	5	5	5
Rank 6	10	6	6	6	9	6	6	6	10	7	6	6	12	7	6	6
Rank 7	9	7	7	6	10	7	7	7	10	9	8	8	14	9	8	8
Rank 8	10	8	7	7	11	8	8	8	13	11	11	10	19	11	10	10
Rank 9	10	8	7	7	10	9	8	8	16	15	14	14	21	15	13	13
Rank 10	10	8	8	8	14	13	13	13	18	19	19	18	23	19	18	18

Table 15: Average Individual-Level Correlation Between Item Rankings and Predicted Rankings by Iterations—Drug Store



	SPARSE				LIMITED				MODERATE				HEAVY			
	100	2000	10000	20000	100	2000	10000	20000	100	2000	10000	20000	100	2000	10000	20000
r =	0.05	0.23	0.25	0.25	0.13	0.46	0.49	0.5	0.42	0.91	0.95	0.95	0.42	0.91	0.95	0.95

Thus far we've shown that:

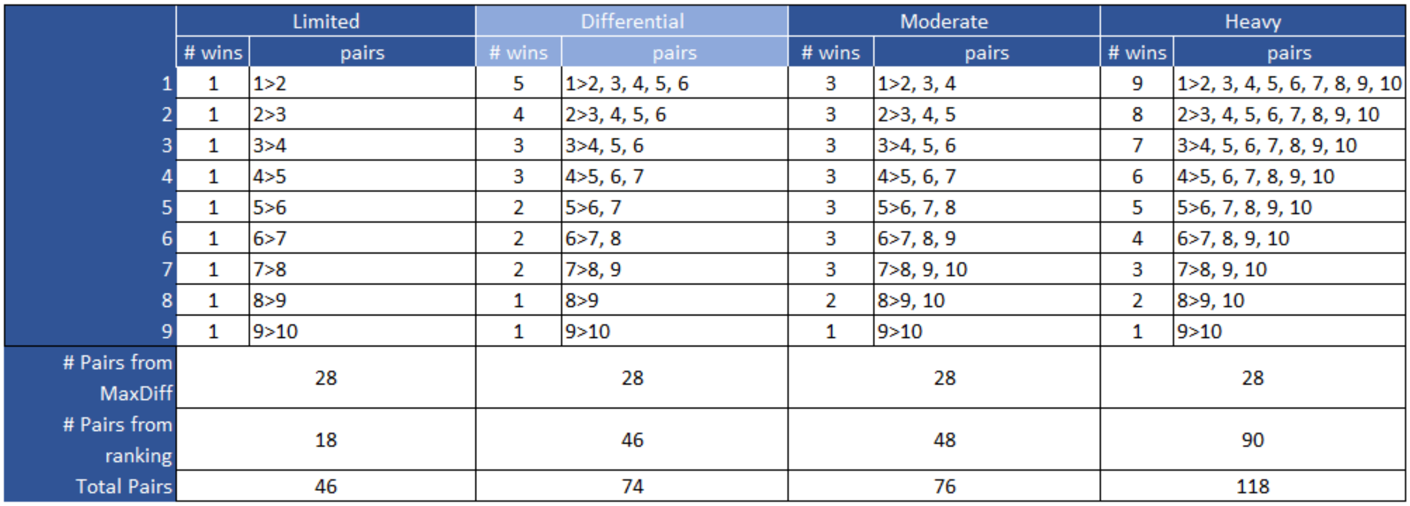
1. Iterations has an independent effect—more iterations yield more extreme values until convergence where it levels out.
2. The point at which convergence happens depends on the level of augmentation—over-specification results in convergence happening after too many extreme values are being predicted.
3. While looking at the percentage of extreme values alone would suggest running fewer iterations, taking a closer look at predictive validity at the individual level based on rank predictions suggests that more iterations are necessary.

Recommendations for Number of Iterations and Level of Augmentation

Don't over-specify the model. Adding too much augmentation causes issues. Somewhere between a limited and a moderate augmentation is the way to go. Looking at the percentage of extreme scores at the point of convergence is a good indication of if you need to increase or decrease the number of augmented pairs going into the design. The levels of augmentation we've discussed thus far all fail in some way—this leads us to ask the question if there is a different way of augmenting that is more carefully considered—Differential augmentation.

We can augment items at the very top more and more (similarly at the bottom) to give more weight to them. In the case of Sparse MaxDiff, the total number of pairs is equal to 28. We then add different numbers of pairs to the design based off the ranking question. Table 16 illustrates how the pairs are specified in Limited, Moderate, Differential, and Heavy augmentation. Limited adds 18 pairs from the ranking, Moderate adds 48 pairs, Differential adds 46 pairs, and Heavy adds 90 pairs. Limited adds all the known pairings of 1 down (1 is better than 2, 2 is better than 3, etc.), Moderate adds all the known pairings of 3 down (1 is better than 2, 1 is better than 3, 1 is better than 4), Heavy adds all possible known pairings (1 is better than 2–10). Differential keeps a similar number of pairs as Moderate, but it specifies the pairings in a non-uniform way keeping more pairings for the top ranked items. This is to give these items more wins in the design.

Table 16: Parings Going into the Design in Drug Store Example



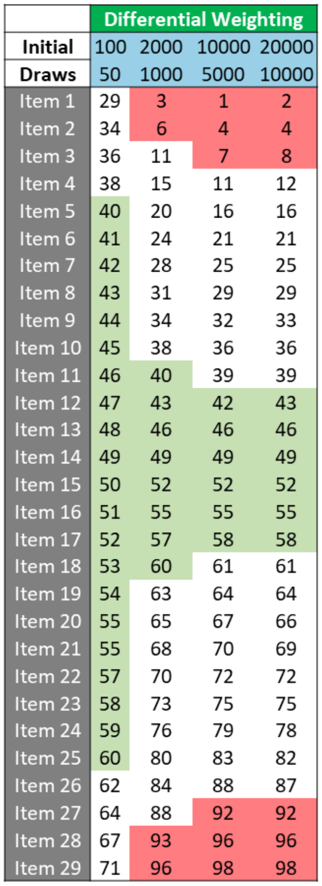
	Limited		Differential		Moderate		Heavy	
	# wins	pairs	# wins	pairs	# wins	pairs	# wins	pairs
1	1	1>2	5	1>2, 3, 4, 5, 6	3	1>2, 3, 4	9	1>2, 3, 4, 5, 6, 7, 8, 9, 10
2	1	2>3	4	2>3, 4, 5, 6	3	2>3, 4, 5	8	2>3, 4, 5, 6, 7, 8, 9, 10
3	1	3>4	3	3>4, 5, 6	3	3>4, 5, 6	7	3>4, 5, 6, 7, 8, 9, 10
4	1	4>5	3	4>5, 6, 7	3	4>5, 6, 7	6	4>5, 6, 7, 8, 9, 10
5	1	5>6	2	5>6, 7	3	5>6, 7, 8	5	5>6, 7, 8, 9, 10
6	1	6>7	2	6>7, 8	3	6>7, 8, 9	4	6>7, 8, 9, 10
7	1	7>8	2	7>8, 9	3	7>8, 9, 10	3	7>8, 9, 10
8	1	8>9	1	8>9	2	8>9, 10	2	8>9, 10
9	1	9>10	1	9>10	1	9>10	1	9>10
# Pairs from MaxDiff	28		28		28		28	
# Pairs from ranking	18		46		48		90	
Total Pairs	46		74		76		118	

Table 17 shows extreme scores for Differential like we showed earlier for Sparse, Limited, Moderate, and Heavy. Differential appears to be converging rather quickly. It has ~20% of items getting extreme scores after convergence, as we'd expect. And, the distribution of items appears to be much more uniform than in other models, with the delta between levels staying roughly the same throughout the list of items.

Table 17: Average Rescaled MaxDiff Utilities for Differential Weighting—Drug Store

Worst item

Best item



	Differential Weighting			
Initial	100	2000	10000	20000
Draws	50	1000	5000	10000
Item 1	29	3	1	2
Item 2	34	6	4	4
Item 3	36	11	7	8
Item 4	38	15	11	12
Item 5	40	20	16	16
Item 6	41	24	21	21
Item 7	42	28	25	25
Item 8	43	31	29	29
Item 9	44	34	32	33
Item 10	45	38	36	36
Item 11	46	40	39	39
Item 12	47	43	42	43
Item 13	48	46	46	46
Item 14	49	49	49	49
Item 15	50	52	52	52
Item 16	51	55	55	55
Item 17	52	57	58	58
Item 18	53	60	61	61
Item 19	54	63	64	64
Item 20	55	65	67	66
Item 21	55	68	70	69
Item 22	57	70	72	72
Item 23	58	73	75	75
Item 24	59	76	79	78
Item 25	60	80	83	82
Item 26	62	84	88	87
Item 27	64	88	92	92
Item 28	67	93	96	96
Item 29	71	96	98	98

Figure 5: A Graphical View of the Distribution with Differential—Drug Store

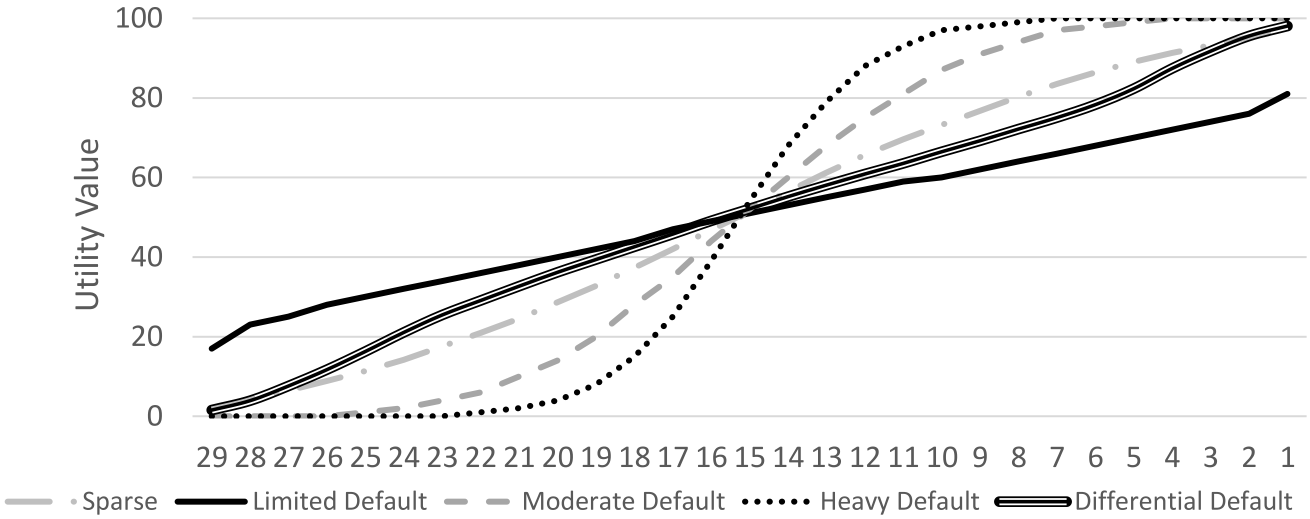


Table 18: Percentage of "Extreme" Scores (<=10, >=90)—Drug Store

	Sparse	Limited	Differential	Moderate	Heavy
Drug Store	**28%**	0%	**21%**	62%	76%
Pain Reliever	50%	11%	**16%**	70%	80%
Laptops	**29%**	0%	**22%**	9%	67%

Figure 5 graphically shows the distribution of utility scores is more in line with what we'd expect. Using Differential weighting gets us in a better range of "extreme" scores as depicted in Table 18. Additionally, we end up with more realistic rank predictions for best and worst ranked items both at the aggregate and individual level. The distribution of rankings based on utilities are also more in line with expectations (Table 21).

Table 19: Median Rankings of Items Selected as Best in the Ranking Question by Augmentation—Drug Store

	Sparse	Limited	Differential	Moderate	Heavy
Rank 1	4	2	1	1	1
Rank 2	5	4	2	2	2
Rank 3	5	5	3	3	3
Rank 4	5	5	4	4	4
Rank 5	5	6	6	5	5
Rank 6	6	6	8	6	6
Rank 7	6	7	9	8	8
Rank 8	8	8	11	10	10
Rank 9	7	8	12	14	13
Rank 10	8	13	14	19	18

Table 20: Average Individual-Level Correlation Between Item Rankings and Predicted Rankings by Iterations—Drug Store

	Sparse	Limited	Differential	Moderate	Heavy
r =	0.25	0.5	0.88	0.95	0.96

Table 21: Distribution of Ranking Based on Utility for Differential Augmentation—Drug Store

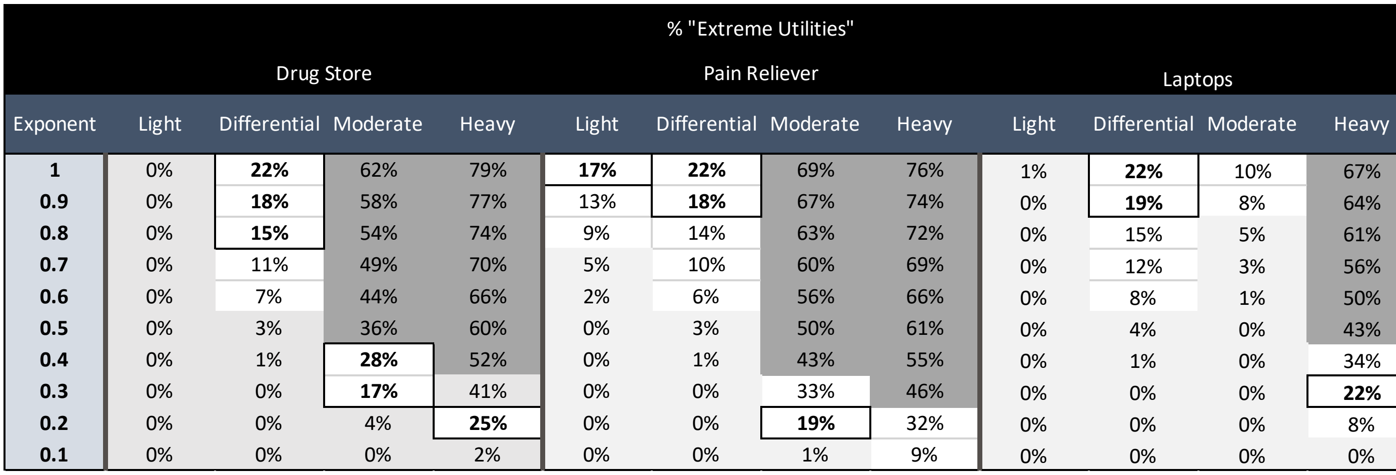
	1st best in rank Q		2nd best in rank Q		10th best in rank Q	
ranking out of all 29 items based on utilities	1	93%	1	7%	<10	18%
	2	7%	2	**88%**	10	**5%**
	3	0%	3	**5%**	11–15	**38%**
	4	0%	4	0%	16–19	**20%**
	5+	0%	5	0%	20+	18%
			6+	0%		

Investigating Other Methods to Control for "Extreme" Scores

One might suggest to fix the "extreme" scores we should modify the exponent in the rescaling procedure. One cause of the extreme scores appears to be a simple scaling issue. Modifying the exponent is designed to correct for this precise situation. In this research, the scaling issue appears to come from a lack of error in our data compared to a standard MaxDiff—where respondents have the freedom to contradict themselves. The ratio of data from the MaxDiff responses to data from the augmentation ranking tasks has a significant impact on the distribution of the utilities. As the proportion of data from the augmentation ranking tasks increases, the extremity of the utilities will increase and an increasingly extreme exponent is necessary. In addition, that change helps with the extreme scores, but has no impact on changes in rank-order preferences.

As mentioned above, while changing the exponent will address the issue of extreme scores, it does not change the rank ordering of those scores. We have a decision to make about number of pairs to go into the design and how to specify the pairs when running Augmented MaxDiffs and we'd prefer to get this right to minimize the amount of post-analysis calibration needed. We did look at the effect of adjusting the exponent and noticed that extreme exponents are needed to correct for the problem. Table 22 shows the impact different exponent factors have on the percentage of "extreme" scores. Sawtooth documentation gives some guidance about exponents specifically for conjoint that we follow even though here we are looking at MaxDiff. The documentation states that: "Exponent adjustments below about 0.2 (for conjoint part-worths estimated via logit-based methods) would seem extreme and point to possible problems in the data (either the part-worth utilities or the holdout judgments being used to tune the exponent)." Here, the requirement of an extreme exponent likely indicates an issue in the number or nature of augmented tasks in the design.

Table 22: Percentage of "Extreme" Utilities



% "Extreme Utilities"												
	Drug Store				Pain Reliever				Laptops			
Exponent	Light	Differential	Moderate	Heavy	Light	Differential	Moderate	Heavy	Light	Differential	Moderate	Heavy
1	0%	**22%**	62%	79%	**17%**	**22%**	69%	76%	1%	**22%**	10%	67%
0.9	0%	**18%**	58%	77%	13%	**18%**	67%	74%	0%	**19%**	8%	64%
0.8	0%	**15%**	54%	74%	9%	14%	63%	72%	0%	15%	5%	61%
0.7	0%	11%	49%	70%	5%	10%	60%	69%	0%	12%	3%	56%
0.6	0%	7%	44%	66%	2%	6%	56%	66%	0%	8%	1%	50%
0.5	0%	3%	36%	60%	0%	3%	50%	61%	0%	4%	0%	43%
0.4	0%	1%	**28%**	52%	0%	1%	43%	55%	0%	1%	0%	34%
0.3	0%	0%	**17%**	41%	0%	0%	33%	46%	0%	0%	0%	**22%**
0.2	0%	0%	4%	**25%**	0%	0%	**19%**	32%	0%	0%	0%	8%
0.1	0%	0%	0%	2%	0%	0%	1%	9%	0%	0%	0%	0%

CONCLUSIONS

Our current research has proven that several clear patterns exist:

1. At the aggregate level, MaxDiff is fully capable of estimating group-level preferences, even with a limited amount of data.
2. Adding augmentation to a Sparse MaxDiff design significantly improves the ability to derive individual-level preferences.
3. Somewhat counter-intuitively, it is possible to include "too much" augmentation.
4. With the correct level and specification of augmentation, utilities will converge, and a reasonable distribution of preferences will emerge.

The need for proper augmentation depends on:

1. The need for accurate individual-level utilities. This is crucial for segmentation, TURF, and other back-end analytics.
2. The level of heterogeneity in the sample: In a more homogeneous sample, the aggregate utilities will reflect the extreme S-curve that we see with too much augmentation, or the relatively flat curve with too little augmentation.

There is more work to be done to perfect the differential augmentation scheme. Additionally, studies can be run to include fixed tasks that we can use to further estimate predictive validity rather than solely referencing the rank order preferences the respondent gave in the follow-up ranking questions.

Jackie Guthart

Curtis Frazier

Raman Saini

References:

Jones, Urzula, and Jing Yeh (2013), "MaxDiff Augmentation: Effort vs. Impact," 2013 Sawtooth Software Conference, Orem, UT. https://www.sawtoothsoftware.com/download/techpap/2013Proceedings.pdf

Sawtooth Software—Lighthouse Studio Help https://www.sawtoothsoftware.com/help/lighthouse-studio/manual/exponentsettingswindow.html

Sawtooth Software MaxDiff System Technical Paper. https://www.sawtoothsoftware.com/download/techpap/maxdifftech.pdf

CAN WE USE RLH TO ASSESS RESPONDENT QUALITY?

MARCO HOOGERBRUGGE
MENNO DE JONG
SKIM

ABSTRACT

RLH is an imperfect indicator of data quality in CBC exercises and it should not be used as a criterion to remove respondents from the data. By means of Latent Class "bad quality" respondents can be identified more accurately.

BEFORE WE START

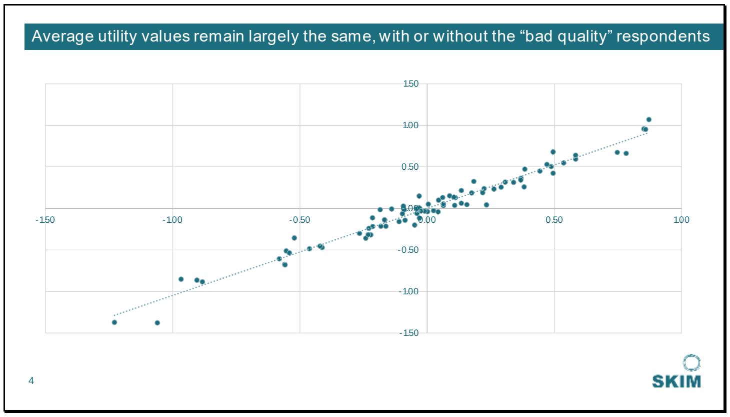
In the research community there is no consensus on whether we should remove "bad quality respondents" at all. Some argue that bad quality respondents may be representative of real people behaving randomly. Also, if we would remove "bad quality respondents," there is no consensus on whether we should remove them directly based on the CBC data, or alternatively just based on more external criteria, such as (page) time spent by respondents. This paper will bypass those more fundamental discussions, however.

The only objection we will look at before we really start is: is data cleaning perhaps just a waste of analysts' time? Would it really change the results if we would keep respondents answering randomly? That is not a trivial question. Taking a non-conjoint example, the *mean score* of a 5-point rating scale will definitely change in case the average score is 4 (because the average of random data is 3). But if the question is part of a grid, and we are merely interested in the *order and relative distances* of the average item scores versus each other, we would not expect a significant impact of random data.

With utilities we are also only interested in relative distances versus each other, so we would not expect substantial differences either. This is confirmed in plots like Figure 1, in which we do not see much difference between average utilities with and without "bad respondents." Note that the definition of "bad respondents" here is based on the procedure that we will elaborate on further down.

Since most of our MaxDiff studies are merely reported in terms of average utilities (with rescaling), this suggests that cleaning out MaxDiff data may not be worth the effort.

Figure 1: Scatterplot of Average Utilities With (x-axis) Versus Without Bad Data (y-axis)



Average utility values remain largely the same, with or without the “bad quality” respondents

With CBC we usually do more than just reporting average utilities, however, because we also simulate the impact of changes in the market. The basis of these simulations are not *average* utilities but *individual* respondent utilities. The reason why simulation results may be impacted by bad data after all is the fact that their HB estimations are guided by the upper-level covariance matrix. For example, there may be positive or negative correlation between utilities of a certain brand and price utilities on an aggregate basis. Random data are price insensitive data (by definition) and by means of the covariance matrix this lack of price sensitivity may be associated more with certain brands and less with other brands. We see an example of this phenomenon in Table 2, in which respondents with random data are often classified in brand B at the cost of brand C.

Table 2: Simulator Shares of “Good” Versus “Bad” Respondents

	Shares of good respondents (91% of original sample)	Shares of bad respondents (9% of original sample)
Premium brand A	27.0%	26.9%
Premium brand B	16.3%	32.4%
Mid tier brand C	24.7%	12.1%
Low tier brand D	11.9%	9.3%
Low tier brand E	16.2%	17.1%
New brand F	3.9%	2.1%

While this phenomenon may not take place in every CBC study, it is definitely a risk and is definitely a reason to check for bad data.

Conceptual Introduction (Illustrated by One Example Study)

This paper is largely based on the idea of adding records with random data to the actual respondents’ records, to then study the *properties* of estimates for the random records and *classify* real respondents as bad respondents if their estimates have the similar properties of the estimates of random records.

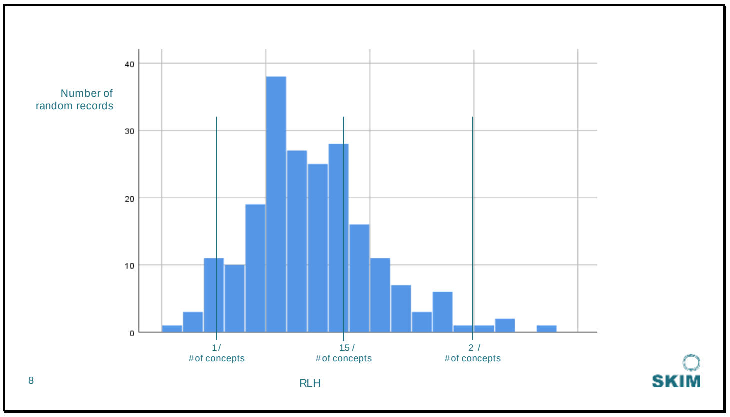
The idea of adding records with random data is not new: it has been published before in https://www.sawtoothsoftware.com/help/lighthouse-studio/manual/index.html?hid_web_maxdiff_badrespondents.html. After adding the random records to the file of the real respondents, HB is run, and we get utilities and RLH values for real respondents and random records simultaneously.

The number of records added to the main data file is always approximately 1/6 of the number of real respondents, throughout this paper. We have determined this ratio in order to have, on the one hand, a large enough number of random records to base conclusions on, while on the other hand the upper model in HB will not too heavily be influenced by adding the random records.

The software help page which was quoted before aims at identifying alternative cut-off values for RLHs to determine bad respondents. The theoretical (average) RLH for random data would be 1/(# of concepts), e.g., 0.25 in case of 4 concepts per screen. In practice a value as low as 0.25 is hardly ever reached because HB will always try and succeed in making some sense of the data, even of random data, so the help page provides a "correction" on the 1/(# of concepts) cut-off, e.g., around 0.32 in case of 4 concepts.

However, as we see in Figure 3, this solution is not quite satisfactory. This example is also based on 4 concepts per task. We see here the confirmation that hardly any respondents are being classified as bad respondents by taking 1/(# of concepts) as a cut-off, hardly any records will be identified as random, but we also see that random data can get an RLH of 1.5 times, or even incidentally of 2 times the minimum cut-off value. While taking the "corrected" cut-off value we would still fail to classify many random records as such.

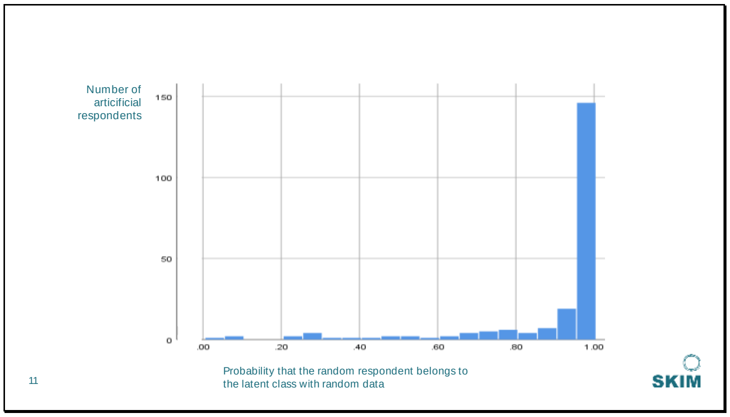
Figure 3: Histogram of RLH Values of Random Data



In addition, we have run Latent Class. Since the amount of random records is about 1/6 of the real respondents, we initially imagined that running Latent Class with 6 groups would be ideal. That way we would have one group with random records (plus bad respondents) and 5 groups with regular (good) respondents.

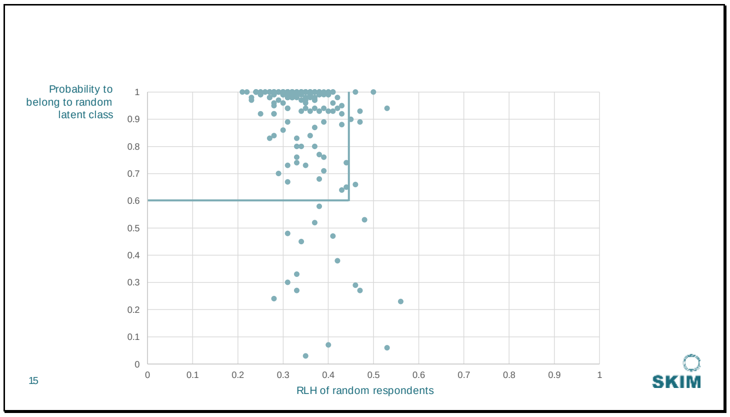
Not surprisingly, the latter hypothesis was confirmed: we had one group with nearly all random records. Figure 4 shows the histogram of probabilities of the random records to belong to that particular group. It is amazing to see that nearly all random records have a probability of more than 0.6 to belong to this group. While it is not 100% perfect (there are still a few random records that have a probability of less than 0.6), this is giving a much more robust confirmation of randomness than the RLH values before.

Figure 4: Histogram of Probabilities of Random Data to Belong to the Random Group



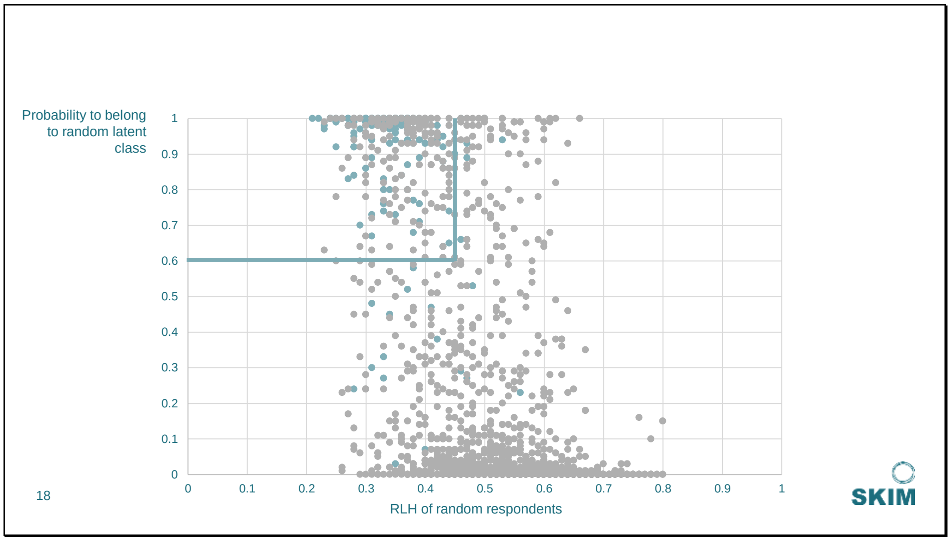
Next, we combined the two previous measures (RLH and probability to belong in the random group) in a scatterplot. For the random records that gives us the picture in Figure 5 where we see that nearly all random records have an RLH below 0.42 and a probability of belonging to the random group of more than 0.6, as depicted by the rectangle.

Figure 4: Scatterplot of RLH and Probabilities to Belong to the Random Group (Random Data Only)



In Figure 5, on the next page, we have added the real respondents and it appears that 9% of the real respondents are in the same rectangle as the random records. We may therefore assume that those real respondents are characterized by having provided random data.

Figure 5: Scatterplot of RLH and Probabilities of Belonging to the Random Group (Real Respondents + Random Data)



Note that in this (simple) case *both* the RLH and the latent class probability contribute to the classification of random data. It may be the case that a real respondent is in the random latent class, but thanks to a decent RLH value they are not classified as a bad respondent. And vice versa. As we will later see, in the end we may even entirely refrain from using the RLH measure.

Finally, in this example, in the raw data we have no fewer than 48 respondents who provide always the same answer in every choice task regardless of the content of that task; they are the so-called flatliners. For example, they always choose the second concept from the left. Theoretically it is possible that these flatlining patterns are just accidental and still driven by real preferences but the chances of that happening are tiny. We may expect that the large amount of them are in fact randomly answering the choice tasks.

This is confirmed in the sense that no less than 38 of our flatliners nicely fall in the rectangle of random records.

GENERALIZATION

While the above introduction was illustrated by one data set, we have experienced that drawing conclusions on just one data set is too risky. In this section we will analyze four different sets, coded as K, L, N, and U. They are relatively similar data sets though: all four contain 5 to 8 attributes and they all four contain 500 to 2000 respondents. It may be tricky to extrapolate the findings below to cases with a very low number of respondents or a very low number of attributes.

We will also take a more structural and efficient approach in the analysis, by using logistic regression models. The dependent variable is whether it is a real respondent or a random record; the independent variables are (various) RLH and LC measures. It should be emphasized that we are *not* directly predicting by what probability a respondent is a bad respondent (we can't!), but rather we predict by what probability a record is a random record and we classify respondents with a high probability as a bad respondent.

EXCURSION: DENSITY VALUE IN UPPER MODEL NORMAL DISTRIBUTION. We have also tried to include another independent variable, namely the density value of respondents' utilities in the HB upper model normal distribution, and non-linear transformations of it. We imagined that a low density value would provide a strong indication that we would *not* be dealing with a random record. However, this did not work out as expected. It was not just a matter of multicollinearity (density and RLH might have correlated too strongly) because even on its own the density value did not predict anything. So it still remains to be seen if we can interpret or use the upper model density value in any meaningful way.

We started the modeling process completely from the beginning, just basing the model on RLH and (for now) ignoring what we have observed in the previous section. Well, in fact we took two measures of RLH, namely the RLH after a "full" HB run of 10,000 burnt iterations and 10,000 saved iterations, and an RLH after a "shortcut" HB run of 1,000 burnt iterations and 1,000 saved iterations. The two measures of RLH correlated hugely with each other, consequently the additional value of a full run in the prediction process is negligible, as shown in Table 6. We were relieved by that, because it implies that we can limit our HB efforts to shortcut runs (at least for the purpose of this paper, for determining an RLH).

Table 6: Fit of Logistic Model Explaining Random Records by RLH

Data set	Pseudo R^2 with RLH after 1K+1K iterations	Pseudo R^2 with RLH after 10K+10K iterations	Pseudo R^2 with both RLHs in the model
K	0.262	0.271	0.271
L	0.273	0.286	0.297
N	0.390	0.339	0.394
U	0.294	0.274	0.297

Note, by the way, that the pseudo R^2 in logistic regression (as introduced by Cox & Snell, 1989) is calculated slightly different than the R^2 in linear regression. Most important to realize is that, in these four studies, the theoretical maximum value of the pseudo R^2 is around 0.6 (rather than 1). So, the model fit is not as bad as it seems at first sight, but it is also far from perfect.

When we now look at Latent Class as a predictor, we can also take multiple Latent Class variables. To start with, we have used Latent Class probabilities from a 6-group solution, and we have taken the Latent Class probabilities of a 20-group solution. An interesting finding here is, while 6 groups were the more "natural" choice of predictors (due to the ratio of adding random records to the data) it appears that, in three of the four studies, 20 groups predicts quite a bit better. In the case of 20 groups we also observed that there were multiple groups that were mainly populated by random records. Apparently, just by chance, they happened to be different enough from each other to form different groups. The obvious recommendation is to increase the number of LC groups as a matter of default data cleaning procedure. In addition, less surprisingly, we found that the additive impact of LC6 and LC20 was quite small compared to LC20 on its own.

Table 7: Fit of Logistic Model Explaining Random Records by LC Solutions

Data set	Pseudo R^2 with 6 segments	Pseudo R^2 with 20 segments	Pseudo R^2 with 6 and 20 segments
K	0.258	0.332	0.327
L	0.198	0.240	0.269
N	0.469	0.497	0.497
U	0.423	0.413	0.433

When we combine the "optimal" RLH (optimal from a combined performance and efficiency perspective) and the "optimal" LC there is only one study with a substantial increase of model fit, compared to taking only the "optimal" LC. In the conceptual introduction the RLH still seemed to play a role in determining random records but once we increase the number of groups in Latent Class significantly, the RLH simply becomes redundant in most cases. For the classification of respondents as "bad quality respondents" we can simply suffice with running a 20-group LC solution (or a high number anyway) on the data set after it has been supplemented with random data.

Table 8: Fit of Logistic Model Explaining Random Records by LC Solutions

Data set	Pseudo R^2 with RLH after 1+1 K iterations	Pseudo R^2 with 20 segments	Pseudo R^2 with RLH1+1K and 20 segments
K	0.262	0.332	0.341
L	0.273	0.240	0.367
N	0.390	0.497	0.499
U	0.294	0.413	0.455

Note that we have tried many more models than described in this paper (especially with alternative LC solutions) but that did not lead to significant new insights.

Conclusion

RLH is a mere moderate indicator of data quality in CBC exercises. Respondents just answering randomly in the CBC tasks can much better be identified by the following procedure:

- Add a substantial number of records with plain random data to the CBC datafile.
- Run Latent Class with a high number (say 20 or 30) of groups.
- Identify which group(s) contain the random records.
- Real respondents in those same groups are the ones that should be removed.

Marco Hoogerbrugge

Menno de Jong

References

Elrod, T. and Chrzan, K. (1994), Partial Profile Conjoint Analysis: a choice-based approach for handling large numbers of attributes. Faculty of Business, University of Alberta, Canada.

Elrod, T. and Chrzan, K. (1994), Choice-based Approach for Large Numbers of Attributes, Marketing News, 29, 20–30.

Green, Paul E and Srinivasan, V (1978). Conjoint Analysis in Consumer Research: Issues and Outlook, *Journal of Consumer Research*, 5(2), 103–123.

Kurz, Peter. A Comparison between Adaptive Choice Based Conjoint, Partial Profile Choice Based Conjoint and Choice Based Conjoint. 2009. SKIM Conference, Prague.

Orme, B. and Johnson, R. (2007) A New Approach to Adaptive CBC. https://sawtoothsoftware.com/download/techpap/acbc10.pdf

BANDIT MAXDIFF: THE EFFECTS OF DESIGN PARAMETERS ON HIT RATES IN DIVERSE DATASETS

NICO PERUZZI
ELUCIDATE

INTRODUCTION

For over a dozen years, MaxDiff has been pushed to study larger and larger numbers of items. In 2007, Hendrix and Drucker used Augmented and Tailored MaxDiff to look at 40 to 60 items. Wirth and Wolfrath (2012) examined upwards of 120 items with Sparse and Express MaxDiff. In 2015, Fairchild, Orme, and Schwartz introduced Bandit MaxDiff as a way to address upwards of 300 items. Finally, Orme (2018) pushed the limits by investigating 1,000 items with Bandit MaxDiff. The desire to explore the use of MaxDiff methods for very large item sets is well established and becoming more and more common in practice.

Sparse MaxDiff emerged as a relatively preferred method for dealing with larger numbers of items (Chrzan, 2015). Sparse and Express MaxDiff both performed well, but Sparse showed a slight edge. However, as the number of items under study continues to increase, Sparse runs into problems of requiring longer questionnaires, more respondents, or both. “Sparse” is well named. Imagine 120 items, showing 5 items per set across 20 sets. With these settings, one is covering all items not even one time per respondent. Still, if one needs information about the bottom (worst) items in addition to the top (best), then one needs to stick with Sparse. However, if the focus is either only on the top (or bottom) items[1], then Bandit has emerged as the most promising technique.

HOW DOES BANDIT MAXDIFF WORK?

Similar to Express MaxDiff, Bandit draws a subset of items (commonly 30) for each respondent. These items become the pool of items used to create the sets seen by each respondent. The difference is that Bandit uses Thompson Sampling to oversample items that are more preferred by the population.

Thompson sampling works as follows: Bandit uses counting analysis “on-the-fly” to determine a best-to-worst order of items for all respondents who have completed the survey up to that moment. These item scores are the percent of wins from exploded paired comparisons. Bandit then perturbs the item scores with normally distributed error with standard deviation equal to the standard error for proportions (SQRT(pq/n)), sorts them, and selects the top items to be included in the subsequent MaxDiff sets.

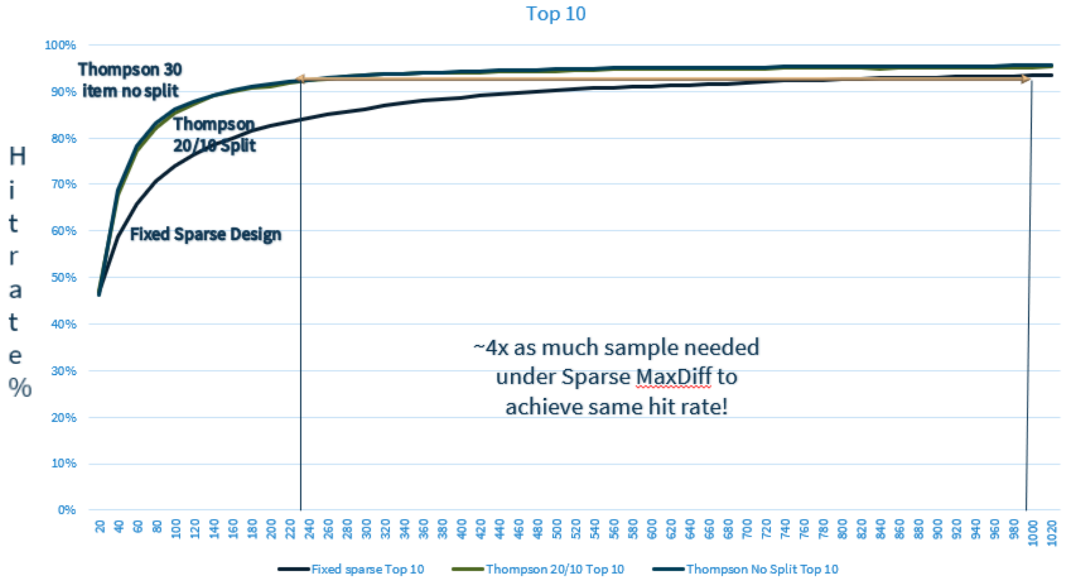
It’s not only the top items that are shown. To protect against non-representative early respondents, Sawtooth Software’s implementation of Bandit, by default, selects some items that have been seen the fewest times by respondents, to give these items a higher probability of being represented in the questionnaire. Importantly, for each new respondent each item

[1] Bandit MaxDiff can be used to prioritize sampling of the worst items and precision in prediction of these worst items if the “best” and “worst” labels are flipped in the questionnaire settings.

has a non-zero chance of being selected as one of the (typically) 30 items to be evaluated. The number of top items included can be varied by the researcher, with the remainder of the items selected among those items seen fewest times to this point by respondents.

The impetus for this current research came after viewing the chart below from the 2015 research of Fairchild, Orme, and Schwartz.

Exhibit 1: Top 10 Hit Rate at Various Sample Sizes for Fixed Sparse vs. Bandit on a 120-item Dataset



That 2015 research showed that Bandit was able to achieve hit rates as good as a fixed Sparse design using approximately one-fourth the sample. Sometimes, sample size isn't the practitioner's biggest concern. Often, concerns regarding survey length and complexity are of equal or greater concern. Thus, the current research focused on what could be done to shorten survey length and reduce survey complexity.

What Are The Parameters One Can Adjust in Bandit, and Why Should We Care?

Before turning to possible parameter adjustments in Sawtooth Software's implementation of Bandit, we considered common differences that may be found in research datasets.

- The number of items studied was investigated, as we questioned whether parameter adjustments would generalize across different sized datasets.
- We also examined the amount of error present in the dataset, as we were curious if extra error in the dataset would affect hit rates, and to what extent. Dataset error can be conceptualized as "noise" and can be viewed as a proxy for sample quality (poorer quality equals more noise and vice versa) and length and complexity of items (more confusing or lengthy item descriptions equals more noise and vice versa).

Turning to the actual parameter adjustments that can be made to Bandit's scripting function, we explored the following:

- The number of items shown to each respondent. We believed that showing fewer items could reduce the cognitive load on respondents, particularly if item descriptions were long or complex. The question was whether such a reduction would negatively affect hit rates.
- The number of items drawn using Thompson sampling (level of adaptivity) deals with what proportion of items likely to be the "stars" according to Bandit's on-the-fly counting analysis are carried forward into subsequent respondents' item sets. (The remaining items are drawn from among those items seen fewest times by previous respondents.) To what degree would more or less adaptivity affect hit rates of the top items?
- The number of screens (sets) shown to respondents. As one of our primary interests was reducing survey length, we explored how far we could reduce exercise length.

To summarize, we studied to following:

- Number of items: 60, 120
- Amount of error in dataset: Standard, High
- Number of items shown per respondent: 20, 30, 40
- Number of Thompson items (level of adaptivity): 5/6, 1/2
- Number of screens (sets) per respondent: 6, 12

The above variables created a total of 48 experimental cells.

THE EXPERIMENTS

We used robotic respondents to simulate results for each experimental condition and conducted 20 replications with unique random seeds for each of the 48 experimental cells. Robotic respondents answered according to true HB utilities plus Gumbel error. The true utilities came from real respondents from the core research dataset from Procter & Gamble on which the original Bandit experiments were conducted (this dataset had 120 items and 984 respondents). All tasks showed 5 items per set.

We created 4 core datasets:

- 120 items standard error: used all items from the original research dataset and randomly selected 300 respondents (see below).
- 120 items high error (double the standard error): used all items but doubled the Gumbel error created in the robotic respondent simulations.
- 60 items standard error: randomly selected 60 items from the original 120.
- 60 items high error: used the same 60 items randomly selected above, but doubled the Gumbel error.

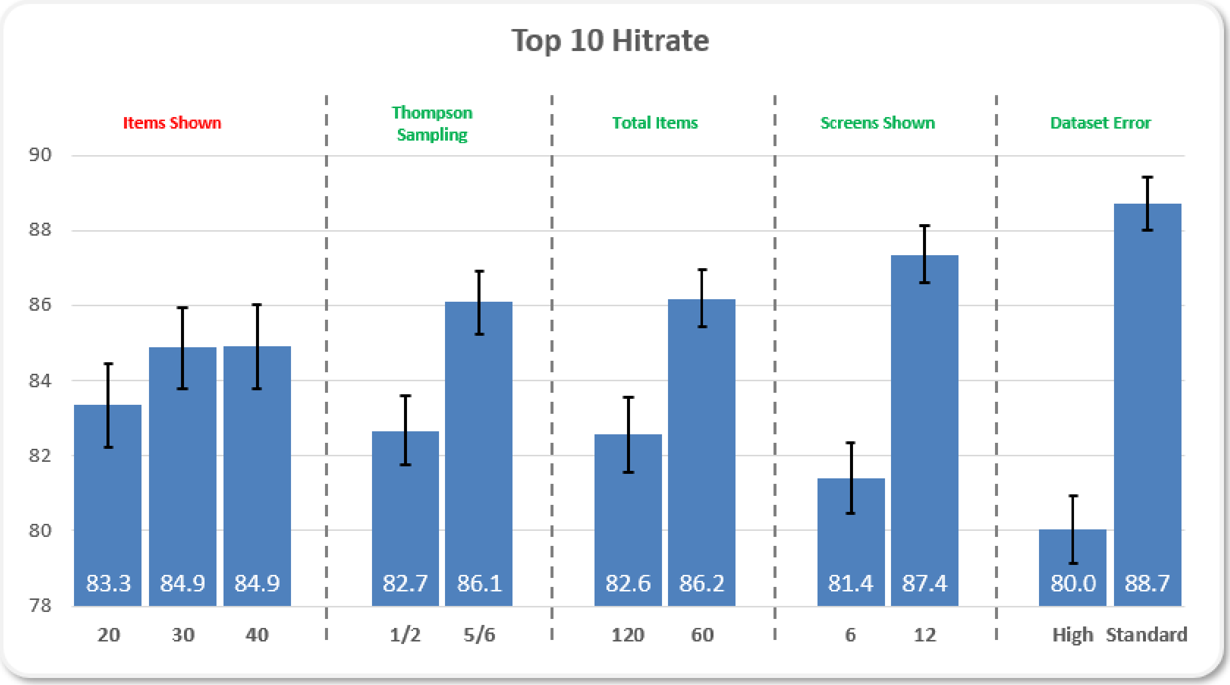
For each of 20 replications per experimental cell, 300 unique respondents were randomly selected and sent through the Bandit MaxDiff exercise. N=300 was chosen as past research showed this sample size providing strong hit rates (given 120 items), with plateauing beyond this number. For each replication, these 300 unique respondents answered according to the true HB utilities plus Gumbel error (to approximate the variation in human behavior).

Hit rates were determined as follows. The actual aggregate top 10 items from the 60 and 120 item true utilities datasets were determined by averaging across the true utilities of all 984 respondents. Top 10 hit rates were defined as the number of top 10 items from each replication in each experimental cell in our experimental trials, as estimated via pooled logit, that matched the true top 10 items in the original dataset. We used a single, simple measure of hit rates to provide an apples-to-apples comparison across conditions where the focus was on relative differences between experimental cells.

Results

In the table below, top 10 hit rates for each experimental variable were collapsed across all other variables. Therefore, N=480 iterations exist across groups for all 2-level groups, and N=320 iterations for the 3-level group (Items Shown).

Exhibit 2: Main Effect Top 10 Hit Rate Comparisons across Levels of Each Experimental Variable



in Top 10. Note: Green indicates 'Significant' differences between comparison groups; Red indicates 'No Significant' differences between comparison groups.

For the Number of Items Shown, interestingly, we found no statistically significant differences in hit rates; though we might have detected a statistically significant difference if we continued with more replications using different random seeds. If the concepts under study are complex, showing respondents a smaller subset of these items (so they have fewer items to orient to) could make their decision process easier to manage. Since we do not lose much in terms of predictive accuracy (hit rate), researchers can consider this option.

For Thompson Sampling, more aggressive adaptivity produces better hit rates in the top 10 items. To clarify the notation in the chart, 1/2 (5/6) means that half (five-sixths) of the items drawn for the respondent use Thompson Sampling and the other half (one-sixth) are drawn from items seen fewest times by previous respondents. As expected, we are better able to home in on top items when we place greater focus on what we are learning from previous respondents (the 5/6 setting). (Though this reduces the precision on items of middling or lower preference.)

For Total Items, of course, the more items we have the harder it is to identify the true top 10, due simply to probability; and with more items, each item is shown fewer times across tasks and respondents leading to worse precision in the logit utility estimates.

For Screens (Sets) Shown, not surprisingly, showing more screens achieves a better hit rate. Although we did not test more than 12 screens in this study, we know from past research that showing 18 screens did not achieve vastly higher hit rates than what we see here for 12 screens. Our goal was to test the lower limit, and there is certainly an effect on hit rate when dropping down to only 6 sets due, again, to the reduction in the number of times each item is seen.

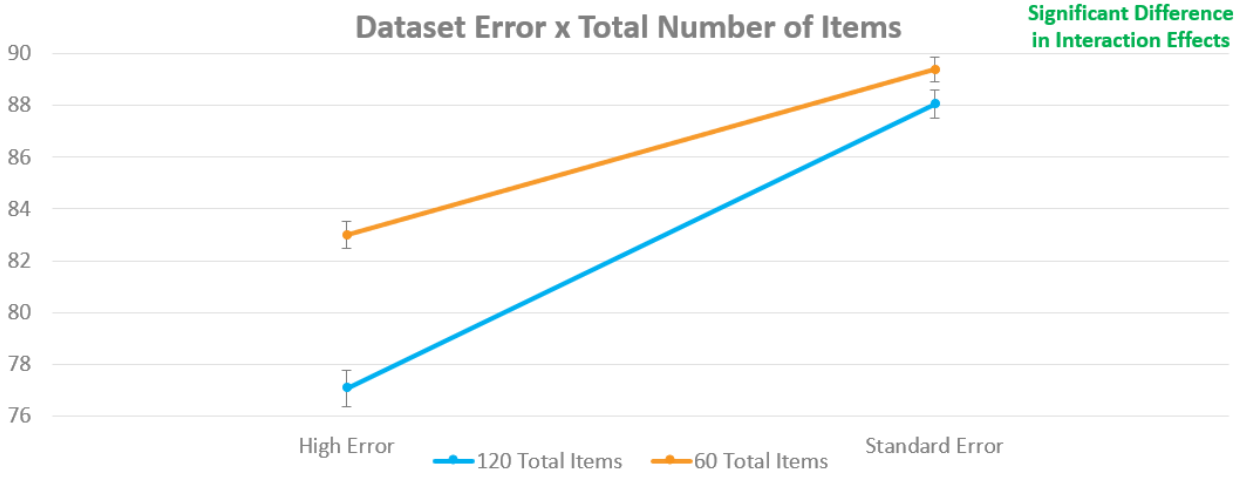
For Dataset Error, as expected, the more error (noise) in the data, the worse the hit rate. If one is working with sample sources expected to be more “dirty” or with item concepts expected to be difficult for respondents to understand, take care to set other parameters such that they will be best able to achieve the highest hit rates possible (i.e., show more screens and use 5/6 Thompson Sampling).

This recommendation holds across all the variables above—a cleaner dataset with fewer items allows for more manipulation of Bandit settings to sub-optimal levels, whereas a dirtier dataset with more items calls for Bandit settings to be set at optimal levels. Note, if additional sample is available, adding more sample (up to a point) will improve hit rates as more respondents will see more items more often, thus improving logit utility estimates.

Few 2-way interactions proved statistically significant, and those that were had limited effects. However, a couple are worth mentioning.

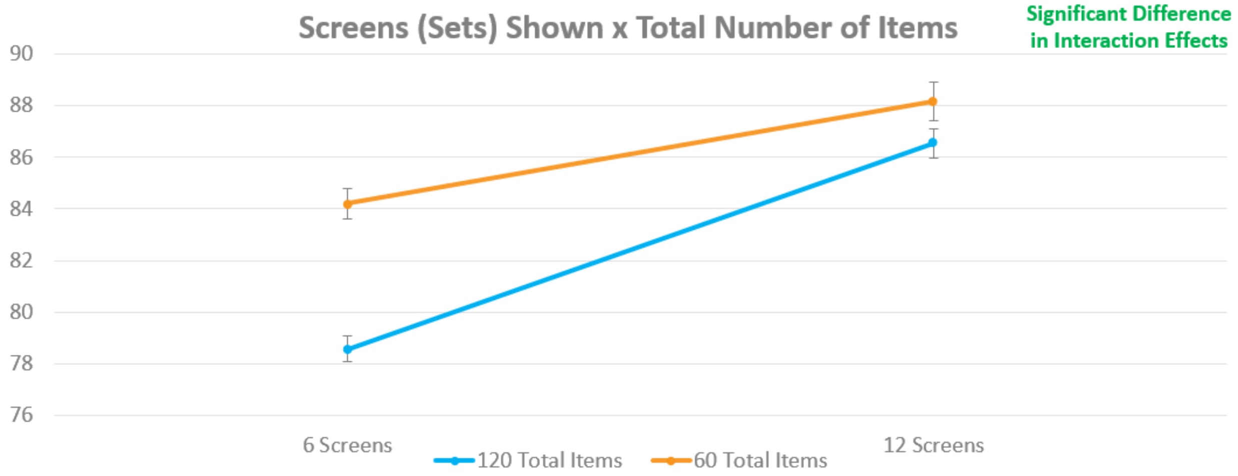
The effect of more dataset error was more pronounced on the larger item dataset (Exhibit 3). Again, if one is working with a dataset expected to have more noise, the larger the total number of items, the more one should keep all adjustable parameters at optimal levels.

Exhibit 3: Interaction between Dataset Error and Total Number of Items on Top 10 Hit Rate



When only showing 6 screens, the drop in hit rate from 60 to 120 total items is more pronounced than the drop seen when showing 12 screens (Exhibit 4). If one is trying to limit the number of sets shown, consider the total number of items. The more items, the more risk you take by reducing sets.

Exhibit 4: Interaction between Screens (Sets) Shown and Total Number of Items on Top 10 Hit Rate



Summary Takeaways

In general, Bandit MaxDiff is fairly robust to all the experimental treatments examined in this research. When approaching any Bandit study, the considerations below provide some guidance.

First, consider the number of items under investigation and how "noisy" a sample source is available. The more items and more sample noise, the more one should optimize parameters to make hit rate success as likely as possible.

Second, consider the length and complexity of item descriptions. The more potential noise caused by item complexity, the more one should optimize parameters to make hit rate success as likely as possible.

In an effort to reduce respondent burden, look for opportunities to decrease the number of sets and the number of unique items shown to each respondent. Number of items shown is the best candidate for reduction without a significant negative impact on hit rate. Reduction in number of sets does affect hit rate, so consider reduction in number of sets in the context of total number of items and potential dataset error. Fewer items and less noise gives more opportunity to reduce number of sets. Again, increasing sample size can help buttress loss in predictive accuracy (i.e., lower hit rate) when adjusting other parameters to sub-optimal levels.

To drive home the above recommendations, consider the following project scenarios:

A 60-item naming study containing short and simple item descriptions. We have a low number of items (low for Bandit, that is) and we anticipate low dataset error due to the simple item descriptions. Reducing the number of items shown to each respondent is less relevant as the items are so simple. Reducing the number of sets will help reduce survey length and will have a limited effect on hit rate. Set Thompson sampling at 5/6, and if available, consider increasing sample size.

A 120-item feature study with long and complex item descriptions. With a larger number of items and greater anticipated error due to long and complex item descriptions, reducing the number of sets shown is less practical. However, the number of items shown to each respondent could be reduced from 30 to 20 in an effort to reduce the cognitive load on respondents. Set Thompson sampling at 5/6, and if available, consider increasing sample size.

A 90-item benefits study with moderately complex item descriptions. Reducing the number of sets could reduce respondent burden, however, at this length and amount of anticipated error, hit rates could be moderately negatively affected. Consider reducing number of sets a small amount (for example to 9 or 10 sets from 12). Reducing the number of items shown to each respondent is possible with limited negative impact on hit rates and could be used to reduce the cognitive load on respondents. Set Thompson sampling at 5/6, and if available, consider increasing sample size.

FUTURE RESEARCH

Does reducing the number of items shown below 20 have an effect on hit rate? How far can we push it down? Taking this idea to an extreme, we need to be concerned about redundancy. As an example, if we only showed 5 items, the set of items shown to respondents would be the same every time.

What else could be explored regarding dataset composition? Does tight vs. disperse clustering in the utilities of the top 10 items affect hit rates? Some studies may have a set of top 10 best items that have utilities very close to each other, whereas other studies may have greater diversity in the utilities of the top 10. Exploring hit rates of the top 3, 5, 7, etc. might reveal different effects of the variables tested in this study.

What about hit rates farther down the item list? Although Bandit is focused on the "top" items, top 10 is a somewhat arbitrary marker. What if there is interest in the "next 10" or some other number farther down the best-to-worst item ranking? Do changes in the level of adaptivity (the Thompson sampling parameter) make a difference in hit rates?

ACKNOWLEDGEMENTS

I would acknowledge the guidance of Bryan Orme throughout this project, the insights and early Bandit MaxDiff work of Kenneth Fairchild, the code-sharing and troubleshooting of Zachary Anderson, and the many hours of help running simulations by my staff at elucidate.

Nico Peruzzi

REFERENCES

Chrzan, Keith (2015), "A Parameter Recovery Experiment for Two Methods of MaxDiff with Many Items," Sawtooth Software Research Paper. Accessed at: https://www.sawtoothsoftware.com/1493

Fairchild, Kenneth, Bryan Orme, and Eric Schwartz (2015), "Bandit Adaptive MaxDiff for Huge Number of Items," 2015 Sawtooth Software Conference, Provo, UT. Accessed at: https://www.sawtoothsoftware.com/download/techpap/2015Proceedings.pdf

Hendrix, Phil, and Stuart Drucker (2007), "Alternative Approaches to MaxDiff with Large Sets of Disparate Items—Augmented and Tailored MaxDiff," 2007 Sawtooth Software Conference, Sequim, WA. Accessed at: https://www.sawtoothsoftware.com/download/techpap/2007Proceedings.pdf

Orme, Bryan (2018), "Bandit MaxDiff: When to Use It and Why It Can Be Better than Standard MaxDiff," Sawtooth Software Research Paper. Accessed at: https://www.sawtoothsoftware.com/1943

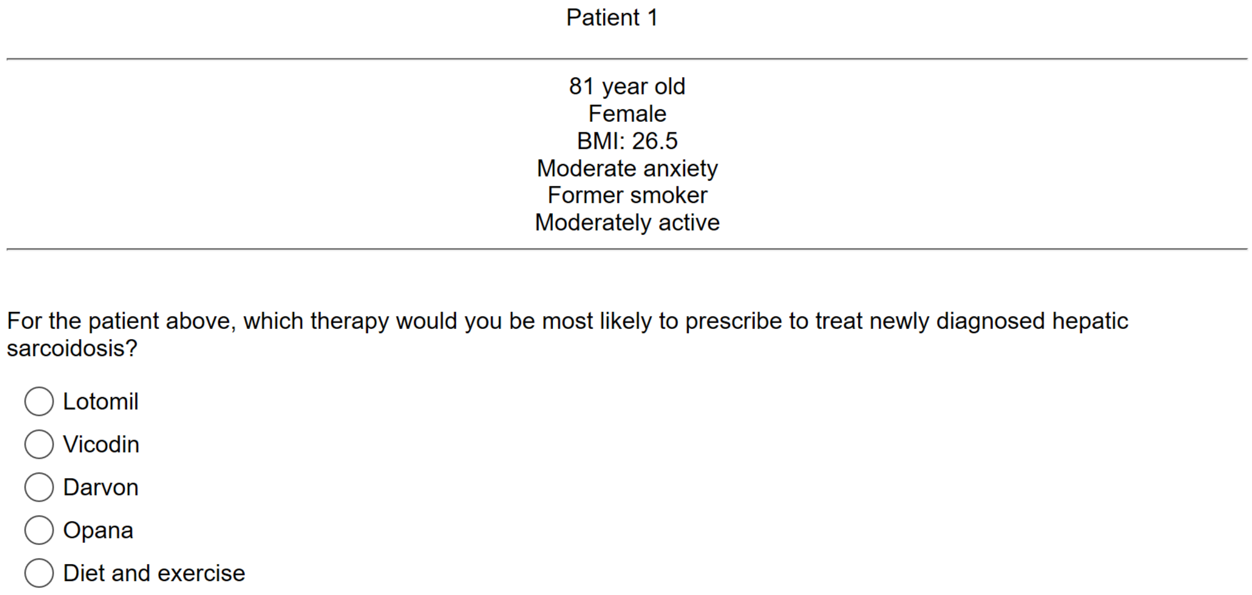
Wirth, Ralph, Anette Wolfrath (2012), "Using MaxDiff to Evaluate Very Large Sets of Items," 2012 Sawtooth Software Conference Proceedings, Provo, UT. Accessed at: https://www.sawtoothsoftware.com/download/techpap/2012Proceedings.pdf

Trees, Forests, and Situational Choice Experiments

Keith Chrzan
Sawtooth Software
Joseph Retzer
ACT-Solutions

Introduction—Situational Choice Experiments

A situational choice experiment (SCE) differs from Choice-Based Conjoint (CBC) experiments in that the experimental design features attributes and levels that describe the choice situation rather than the choice alternatives. In other words, the attributes and levels are invariant across the choice alternatives. A common example occurs in pharmaceutical marketing research, where we might create a set of experimentally designed patients, defined in terms of therapy-relevant attributes such as demographics (age, sex), their disease state (therapy history, progression, stage), concomitant conditions, and insurance coverage. We might then have physician respondents choose which of several therapies they would prescribe to each of the patients described in the experimental design. For example, here is a disguised example of one SCE question from a recent study for a pharmaceutical client:



Patient 1

81 year old
Female
BMI: 26.5
Moderate anxiety
Former smoker
Moderately active

For the patient above, which therapy would you be most likely to prescribe to treat newly diagnosed hepatic sarcoidosis?

○ Lotomil
○ Vicodin
○ Darvon
○ Opana
○ Diet and exercise

As in a CBC, we would ask several of these questions and subsequent questions would describe different patients (e.g., an inactive, non-smoking 62-year-old man with a BMI of 30.0 and moderate anxiety).

Other examples of SCEs include (a) modeling retirement decisions among end-of-career workers as a function of economic factors, (b) modeling the purchase/rent/neither decision about an expensive industrial durable as a function of the attributes and levels of that product, or (c) modeling consumers' preferences among compensation options for service failures, as a function of attributes and levels describing those failures.

Rather than the conditional multinomial logit (MNL) used on CBC experiments, which produces a single vector of utilities from a set of profiles that describe the different choice alternatives, we can use polytomous (or unconditional) MNL to estimate utilities. With each of the several choice tasks including just one experimentally designed profile that is invariant across choice alternatives, the polytomous logit produces alternative-specific sets of utilities; in the example above, the polytomous MNL model would produce utilities for each level of each attribute, plus an alternative specific constant, for each of the five choice alternatives (Hosmer and Lemeshow, 2000; Hoffman and Duncan, 1988). For example, the utilities from the experiment above might look like this:

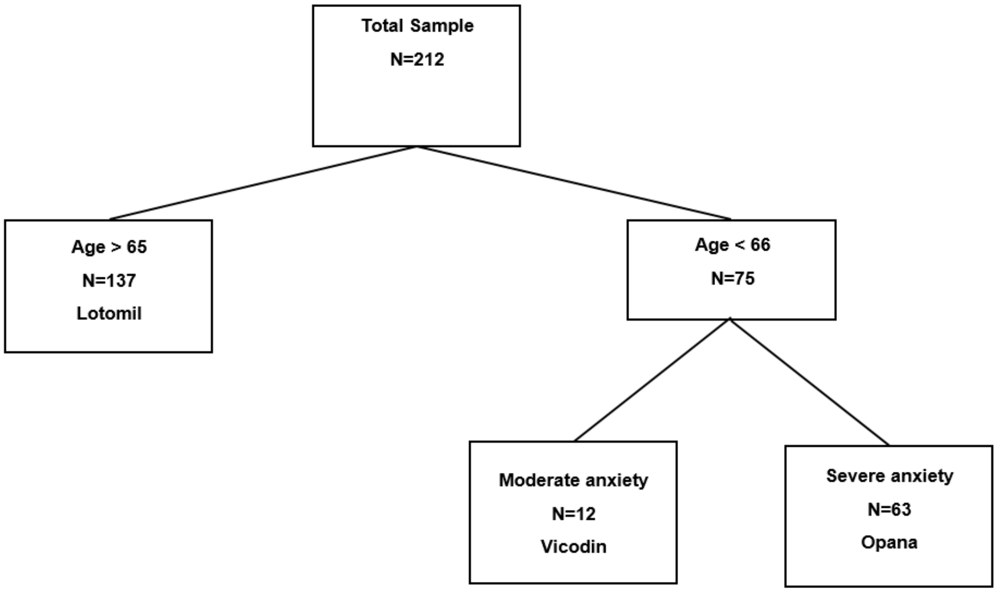
Attribute Level	Lotomil	Vicodin	Darvon	Opana	Diet and Exercise
25 year old	0.42	1.07	-0.29	-0.45	0
43 year old	-0.46	-0.42	0.84	0.40	0
62 year old	1.15	-0.56	0.79	0.50	0
81 year old	0.36	1.18	-0.83	-0.15	0
88 year old	0	0	0	0	0
. . .	. . .	. . .	. . .	. . .	. . .
No anxiety	-1.60	-0.25	1.27	-0.02	0
Moderate anxiety	-0.67	-0.13	-0.93	-0.71	0
Severe anxiety	0	0	0	0	0

Note that the Diet and Exercise column has all zeros because it serves as the reference level that allows the model to be identified.

As we do with CBC, we can use these utilities to build a simulator so that our clients can run sensitivity analyses and see how the share results change across different patient profiles. More so than for CBC, we encourage clients to interpret the model using a simulator rather than any kind of narration of the utilities, for reasons the complexity of the model makes clear.

Machine Learning Modeling Options

While it’s natural for choice modelers to think about using logit to estimate utility functions, several machine learning methods may also apply. For example, Classification and Regression Trees (CART) can predict categorical variables like a brand choice (Breiman et al., 1984). Decision tree analyses like CART usually come up when a client wants segmentation (e.g., a patient type segmentation in the example above). We can also use trees for prediction. Visualization provided by the tree may be easier for some clients to absorb than a polytomous MNL model with five vectors of utilities as above. For example, for the disguised study above, the tree might look something like this:



In addition to trees, we might think about analyzing our SCE with random forests. As the name suggests, random forests analysis doesn't give us a single tree, but rather a large number of trees, each with some random perturbations (random subsets of respondents, random subsets of predictors at each node). The random perturbations effectively "de-correlate" the trees so that the forest doesn't suffer from the collinearity problem that can affect individual trees (while an SCE uses a designed experiment to create the profiles, the collinearity problem can be severe when modeling observational or cross-sectional data). Random forests analysis enables predictions and it provides a measure of attribute importance we could report to our clients (Breiman, 2001). We can construct forests from different kinds of tree models, so in addition to building a forest from CART trees, we could build one from conditional inference trees—a specific variety of tree model designed to avoid favoring splits on continuous attributes or attributes with otherwise large numbers of levels (Hothorn et al., 2006).

Another possibility is boosting, which involves building a succession of trees where each tree models the prediction errors of the previous tree. Individual trees each add a bit to the prediction of the overall model. A variety of hyperparameters control the model estimation process. The first boosting model we will consider is "eXtreme Gradient Boosting" (XGBoost).

XGBoost is a fast, regularized implementation of gradient boosting. Regularization is a technique employed to combat a common problem in boosting models known as overfitting. Overfitting happens when our model conforms very closely to the data used to build the model and consequently does a poor job of predicting new observations. Regularization makes XGBoost a more robust and accurate variant of gradient tree boosting.

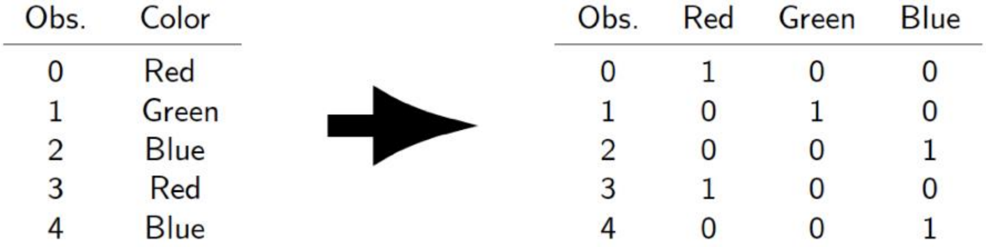
The "eXtreme" in "eXtreme Gradient Boosting" refers to the engineering involved in pushing computational resource limits for the model through:

- parallelizing split finding with each tree,

- pre-analysis transformations of the data, and
- utilizing partially compiled code,

as well as a variety of additional enhancements causing XGBoost to train models quickly. It also enables the model to effectively scale up to large data sets.

XGBoost requires categorical features to be dummy coded using the "One-Hot-Encoding" scheme. This approach replaces each level of the categorical feature with its own column of 1s and 0s as illustrated below:



Obs.	Color
0	Red
1	Green
2	Blue
3	Red
4	Blue

Obs.	Red	Green	Blue
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	0	1

It's clear from the above illustration that "One-Hot-Encoding" may result in a substantial increase in data matrix dimensionality, particularly when our categorical features are of high cardinality (many levels). While datasets used in this study did not include categorical features of unusually high cardinality, it was the case that ALL features in every model were categorical. Since many features with average cardinality may also result in an undesirable increase in data dimensionality, it was decided to employ an alternative approach, CatBoost, a boosting algorithm which provides native support of categorical features.

CatBoost is a relatively new, open source, machine learning library based on boosted gradient decision trees, created by a Russian company called Yandex (Dorogush et al., 2017). It offers native support for both numerical and categorical features. The model addresses overfitting through something called "ordered boosting." Basically, ordered boosting involves permuting all observations in the data set and then predicting the value of the categorical variable in question using only the previous observations in the ordering.

While CatBoost may use "One-Hot-Encoding" for categorical features, it can also re-code categorical features directly using an approach referred to as "ordered target statistic encoding." Again, briefly, this entails permuting the data and replacing the levels of the categorical variable with a number. The number is calculated using previous cases in the permutation.[1]

This prevents an increase in data set dimensionality and hence CatBoost was chosen as an alternative to the XGBoost model.

CatBoost boasts superior prediction speed utilizing GPU processing power and comparatively stable hyper-parameters which therefore requires less model tuning. CatBoost proved to be significantly slower to train however, making model exploration more difficult and time-consuming.

[1] For a more detailed description of both "ordered boosting" and "ordered target statistic encoding" see "CatBoost: unbiased boosting with categorical features" by Dorogush et al.)

Previous Research

Bock (2019) compares the results of using a decision tree and a logistic regression to analyze data from an experiment with a binary dependent variable. Finding that the decision tree predicted the binary choice better than did the logit model, Bock also notes that the decision tree enabled a better story to communicate the meaning of the model.

Brathwaite et al. (2017) note that because they can uncover non-compensatory choice processes, decision trees make good sense from a micro-economic standpoint. Using a cross-sectional (i.e., non-experimentally designed) data set, Brathwaite et al. found that a Bayesian decision tree predicted consumers' choices better than did a polytomous MNL model.

Again using cross-sectional data (respondent-reported travel mode choices) to build a revealed preference model, Sekhar et al. (2016) found that a random forest model predicted those choices better than did a MNL model.

Finally, Lhéritier, et al. (2017) used observed behavioral data (flight bookings) and again found that random forests predicted fliers' choices more accurately than did MNL.

Our review of the literature comparing logit choice models with trees and forests suggests the superiority of the latter.

Empirical Study for Predictive Validity

The first of several comparisons we want to make of logit models, decision trees, random forests, and CatBoost involves predictive validity.

Data

The analyses below use 10 empirical data sets. The first nine in the table below come from situational choice experiments in which physician respondents made decisions based on experimentally designed patient profiles. The tenth study used data drawn from patient records, so the patient profiles lacked experimental control, but the patient profiles and therapy choices were real, not hypothetical.

Study	MNL parameters	Respondents	Tasks/ Respondent	Alternatives/ Task
1	75	313	7	6
2	38	70	6	3
3	40	95	14	3
4	26	253	12	3
5	39	320	15	4
6	36	710	7	5
7	102	600	13	7
8	42	110	16	3
9	84	400	8	8
10	22	286	5	3

The large number of parameters in each model owes to the fact that each choice alternative has its own alternative-specific set of utilities. Of course, the parameter count applies only to the logit model, but it effectively communicates how complex those models can be. The largest study had 13 x 600 = 7,800 observations while the smallest had just 420 observations.

Models

We built the following choice models for each of our 10 data sets:

- Polytomous MNL
- CART decision tree
- Random forest of CART trees (RF)
- Random forest of conditional inference trees (cforest)
- CatBoost

Model Training and Evaluation

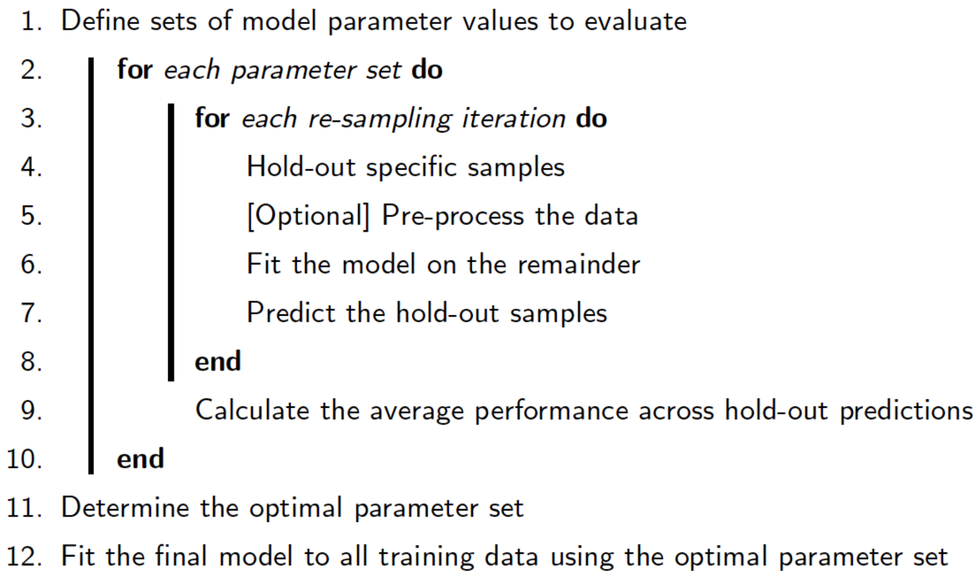
Training the CatBoost model involves identifying optimal values of parameters that control model estimation, aka “hyperparameters.” The primary CatBoost hyperparameters include:

- Maximum tree depth
- Leaf estimation iterations
- Learning rate
- Iterations total
- Border count (number of splits considered for each feature)

A grid search across all combinations of hyperparameter levels was performed using the caret package in R. The caret package begins by selecting a single set of hyperparameter levels from a matrix composed of all combinations of specified levels (this matrix is generated by caret as well).

With a selected set of hyperparameters, caret performs n-fold (in this case n=10) cross validation, averaging predictive accuracy across the estimated models. The final average provides a measure of accuracy for the CatBoost model based on the specific hyperparameter value set selected. Caret then selects another set of hyperparameter levels and repeats the process. A comparison of the of accuracy for each model (one for each unique combination of hyperparameter levels) is used to select the optimal model. The optimal model is then applied to the "test" data to get a measure of accuracy based on data that was not involved in the model estimation.

An illustration of the caret model hyperparameter search algorithm described above is given below:



```
Define sets of model parameter values to evaluate
    for each parameter set do
        for each re-sampling iteration do
            Hold-out specific samples
            [Optional] Pre-process the data
            Fit the model on the remainder
            Predict the hold-out samples
        end
        Calculate the average performance across hold-out predictions
    end
Determine the optimal parameter set
Fit the final model to all training data using the optimal parameter set
```

Results

The test of predictive validity (hit rates) produces clear results: CatBoost predicts better than the other models in every one of the 10 studies:

Study	MNL	CART	RF	cforest	CatBoost
1	44.5	42.9	41.7	41.4	45.2
2	60.0	59.3	58.8	59.0	61.4
3	64.0	66.2	62.3	64.7	72.5
4	57.7	57.8	56.8	57.2	59.1
5	49.7	49.4	48.4	48.8	51.1
6	48.1	48.4	47.6	48.2	49.7
7	35.5	35.1	35.0	34.7	36.5
8	51.3	50.8	48.9	51.0	51.7
9	41.7	42.2	42.3	41.9	46.7
10	61.5	61.1	63.1	62.6	63.4

Our results suggest the computational cost (the additional time spent training the CatBoost model) was worthwhile as its predictive accuracy consistently proved best.

After CatBoost, MNL predicts better than CART in 6 of the 10 studies, and better than a forest in 6 of the 10 studies. Thus, we are unable to replicate the results we found in the literature review about the superiority of trees and forests over MNL for choice modeling applications. Nor did our studies show the predictive superiority of forests over trees (in fact, the CART decision tree outperforms both random forest models, in terms of prediction, in 7 of the 10 studies). We suspect this latter finding may owe to the nature of our data. Our experimentally designed profiles in studies 1–9 prevent collinearity among predictor variables, so methods like random forests that use sample and variable randomizations to counteract collinearity just don't have the advantage they might in a cross-sectional study (and in the only cross-sectional data set we have, study 10, random forest methods do outperform both the CART decision tree and the polytomous MNL model).

Evaluation

In terms of prediction, CatBoost performed better than all the other models. If your only goal for a SCE is in predicting choices, then CatBoost appears to be the way to go.

Often clients want not only to predict choices, but to understand them as well, so explanation can also be important. CatBoost is something of a black box when it comes to explanation. Random forests provide importance scores, which, while they don't map directly to choices, help explain which attributes drive the choices. Polytomous MNL, with its coefficients, standard errors, and odds ratios, lends itself to explanation and inference. Decision trees also explain choices, via a series of if-then dichotomies.

Decision trees have a further advantage in that they allow us to visualize the respondents' decision process. Clients often speak in terms of wanting to understand "decision hierarchies" and the graphical output of a CART analysis provides exactly such a hierarchy.

Finally, clients typically want an Excel simulator as a deliverable for their SCE study. Polytomous MNL provides logit-based simulations familiar to choice modelers. Decision trees, simply using a set of if-then rules, also work well when brought to life in Excel simulators. Simulations are possible with random forests; though these are easy to do in R (simply run a new data set through the forest built from your experimentally designed data and give each tree one "vote"), building one in Excel would be a formidable feat. Like a random forest, CatBoost would be difficult to build into an Excel reporting tool.

Recommendation

If your only objective is accurate prediction, use CatBoost. If you need to understand what drives respondents' choices, use decision trees or polytomous MNL —both methods show how attributes and attribute levels relate to choices, the former in a more statistical way and the latter in a way more amenable to visualization. Also consider decision trees if your client mentions wanting to understand a decision hierarchy.

Keith Chrzan

Joseph Retzer

References

Bock, T. (2019) "Decision Trees are (Usually) Better Than Logistic Regression," *Quirk's Marketing Research Review*, January, 52–55.

Brathwaite, T., A. Vij and J.L. Walker (2017) "Machine Learning Meets Microeconomics: The Case of Decision Trees and Discrete Choice," arXiv preprint arXiv:1711.04826.

Breiman, L., J.H. Friedman, R. A. Olshen and C. J. Stone (1984) *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Breiman, L. (2001) "Random Forests," *Machine Learning*, **45**: 5–32.

Dorogush, A.V., A. Gulin, G. Gusev, L. Ostroumova Prokhorenkova, and A. Vorobev. (2017) Catboost: Unbiased Boosting with Categorical Features. arXiv preprint arXiv:1706.09516.

Hoffman, S.D. and Duncan, G.J. (1988). "Multinomial and Conditional Logit Discrete-Choice Models in Demography," *Demography*, **25**: 415–27.

Hosmer D.W., Lemeshow S. (2000). *Applied Logistic Regression*. New York: John Wiley and Sons.

Hothorn, T., K. Hornik and A. Zeileis (2006) "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, **15**: 651–674.

Lhéritier, A., M. Mocamazo, T. Delahane and R. Acuna-Agost (2017) "Airline Itinerary Choice Modeling Using Machine Learning," (2017), http://dx.doi.oerg/10.1016/j.jocm.2018.02.002.

Sekhar, C.R., Minal, and E. Madh (2016) "Mode Choice Analysis Using Random Forrest Decision Trees," *Transportation Research Procedia,* **17**: 644–652.

EXAMINING THE NO-CHOICE OPTION IN CONJOINT ANALYSIS

MAGGIE CHWALEK
ROGER A. BAILEY
GREG M. ALLENBY
OHIO STATE UNIVERSITY

ABSTRACT

For valid economic interpretation a conjoint analysis must include, at a minimum, each alternative's brand name, prices, and an outside "no-choice" option. Brand names serve as proxies for the attributes not mentioned in the study, and prices allow for economic calculations. The no-choice option is offered in case respondents determine that some other offering, not included in the choice set, is preferred to those included in the choice set, or that it is better for them to hold on to their money and not make a purchase. Thus, selecting the no-choice option assumes that respondents have some knowledge of the prices and features of the real marketplace. In this paper, we examine the effect of providing respondents with information about the prices and features of real category options on conjoint preference estimates. Using a national sample of respondents in the tooth whitening category, we find that conjoint estimates are surprisingly robust to the information provided about existing marketplace options.

1 INTRODUCTION

Conjoint analysis relies on an economic view of choice where respondents are able to recall and construct their preferences for hypothetical offerings (Ben-Akiva et al., 2018; Manski et al., 1981). The utility part-worths for product attributes and their levels are assumed to be recalled independently for each product feature and combined with other part-worths to arrive at an overall measure of utility that determines preferences. The independence assumption is needed to rationalize choices in terms of part-worth utilities and provides a basis for using different respondent choices to obtain a common set of part-worth estimates. Interactions among attribute-levels are accommodated by including appropriate terms in the model specification, with the assumption that these interactive effects can also be independently recalled by respondents (c.f., Lichtenstein and Slovic (2006)).

The minimal requirements for conducting a conjoint study to study marketplace demand is the inclusion of product brand and price. Brand is needed for respondents to imagine the offering and recall associated attributes and levels of performance. The brand name serves as a proxy for the unmentioned attributes of a brand in a conjoint study, and consumer knowledge of these features is what gives the brand names its value. Price is required because marketplace transactions involve prices that are also needed to compute measures of economic value, such as willingness to pay (WTP) and price premiums (Allenby et al., 2014b). A conjoint analysis with only brand and price is often referred to as a brand-price tradeoff study (Rao, 2014). However, conjoint studies usually include other product attributes or features of interest to the analyst. Attributes that are familiar to and understood by respondents factor into their choices, but those that are unfamiliar or not understood are less important in the decision process (Balbontin et al., 2017; Sandorf et al., 2017).

The "no-choice" option in a conjoint study allows respondents to choose something other than one of the brands included in the choice task (Brazell et al., 2006). Not only does the no-choice option serve to increase the realism of the respondent's decision, it can improve the resulting market share and volume predictions from the analysis (Carson et al., 1994). By selecting the no-choice option, respondents are indicating preference for an outside option, or the desire to opt out of the market altogether and save their money. This allows respondents to compare the utility of the alternatives in each choice task to some fixed level of utility they know they could achieve outside of the market (Louviere et al., 2010; Bahamonde-Birke et al., 2017). This interpretation of the outside good choice is valid only if consumers are familiar with the attributes and prices of the outside options left out of the survey.

Consider, for example, the price of assisted living care for the elderly. It is doubtful that individuals would be aware of the daily rate of assisted care unless they were involved in the financing or arrangement of care for an elderly person. Respondents probably have a much better grasp on the incremental value of increased services, such as improved dining options. Since the selection of the no-choice depends on the overall price level of outside options, there is a chance that it is selected at a rate that is not consistent with how respondents actually act in the market where they are spending their own money for goods and services.

Surveys routinely screen for respondents who are "in" the market using questions about purchase intent, purchase history, and participation in the buying decision (Ben-Akiva et al., 2018). In addition to assuring familiarity with the product attributes, screening questions provide some degree of assurance that consumers are viable prospects in that they are expressing their willingness and ability to make purchases in a product category. Moreover, screened participants have a higher likelihood of being aware of general price levels because of their past or intended purchase activities.

The purpose of this paper is to examine respondent sensitivity in a conjoint study to information about marketplace prices and product features. If standard screening questions about product participation are sufficient for identifying qualified candidates who are aware of marketplace prices, then providing additional pricing and attribute information will minimally affect part-worth estimates. However, if survey participation questions are an insufficient proxy for brand price and attribute knowledge, then providing price and attribute information will change consumer preferences for the outside good and part-worth estimates in general. Invariance of consumer choices to outside good information is therefore fundamentally related to the assumption that the economic view of preference construction is valid for conjoint studies with properly screened respondents.

The organization of this paper is as follows. Section 2 provides a description of the conjoint survey including screening questions and the experimental stimuli used to test the robustness of the part-worth estimates. Section 3 summarizes the data and model parameter estimates based on a standard hierarchical Bayesian random-effects model. A comparison of results across information sets is provided in Section 4. Section 5 provides concluding comments.

2 Tooth-Whitening Study

We investigate the effect of outside good information on respondents' choices using a conjoint study of tooth-whitening products. Participants were recruited from a national panel for participation. The tooth-whitening category was selected because products exhibit large price variation and a variety of product attributes. A list of product attributes used in the survey are presented in Table 1 along with their definitions.

Table 1: Products Attributes, Definitions, and Levels

Attribute	Definition	Levels							
		1	2	3	4	5	6	7	8
Brand	The brand name of the teethwhitening product	Crest	PluWhite	Rembrandt	GoSmile	Auraglow	Luster		
Form	The method of application of the product	Strips	Pen	Trays + Gel	Light Tech				
Treatment Time	The total suggested time for a single whitening treatment	5	15	25					
Number of Treatments	The total suggested number of whitening treatments to be completed	7	14	21					
Time Until Results	The claimed amount of time until consumers should see visibly whiter teeth	3	7	14					
Peroxide %	The percentage of the active ingredient (hydrogen peroxide) contained in the product	6	7	8	9	10	11	12	13
Price	The price of the product	14.99	24.99	34.99	44.99	54.99			

Individuals responding to the Internet panel provider's invitation to participate were presented with a series of screening questions to determine whether the potential respondent was "in" the product category and had sufficient knowledge to provide informative answers to the survey questions. To be included in the survey, respondents must make their own hygiene purchases, be medically qualified to use the product, and be engaged in or actively considering buying some offering in the product category. The screening questions used in the survey are presented in Table 2.

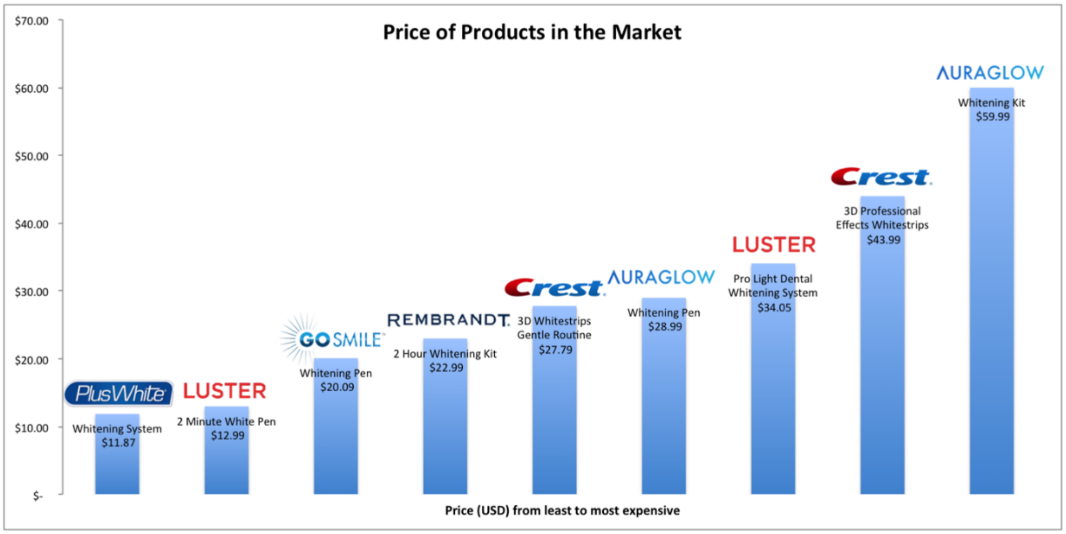
Table 2: Screening Questions

Which of the following describes your level of involvement in purchasing hygienic products (soap, shampoo, toothpaste, etc.) in your household? ○ I do all of the purchasing of hygienic products in my household ○ I share the responsibility of purchasing hygienic products in my household ○ I am not involved in purchasing hygienic products in my household	Please select all of the following dental/orthodontic work you CURRENTLY have or PLAN to have done in the next three months. ○ Braces ○ Dentures ○ Caps, crowns, veneers, or fillings—not visible when smiling (back teeth) ○ Caps, crowns, veneers, or fillings—visible when smiling (front teeth)
Have you purchased a teeth whitening product in the past? ○ Yes ○ No	Would you consider buying a teeth whitening product? ○ Yes ○ No

Qualified respondents who successfully passed the screening questions were then required to watch a short video describing the choice task and defining the product attributes and their levels.[1] Respondents were then randomly assigned to one of three experimental conditions. The first condition did not provide any information about marketplace prices or attributes and serves as a control group for the other two conditions.

The second experimental condition provided respondents with the range of prices for the different brands under study as well as brands not included in the survey. Prices were obtained from the posted prices on Amazon.com. Figure 1 displays the information provided to respondents.

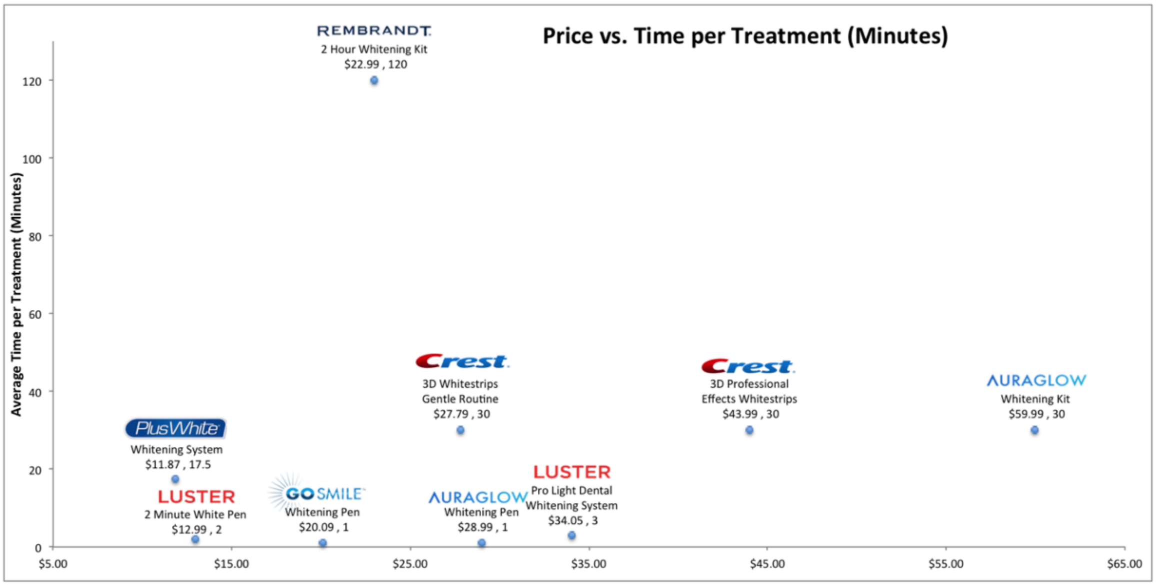
Figure 1: Price Distribution



[1] The video can be viewed at https://mediasite.osu.edu/Mediasite/Play/947f2aa2ec284abbb1166d66df86d8711d

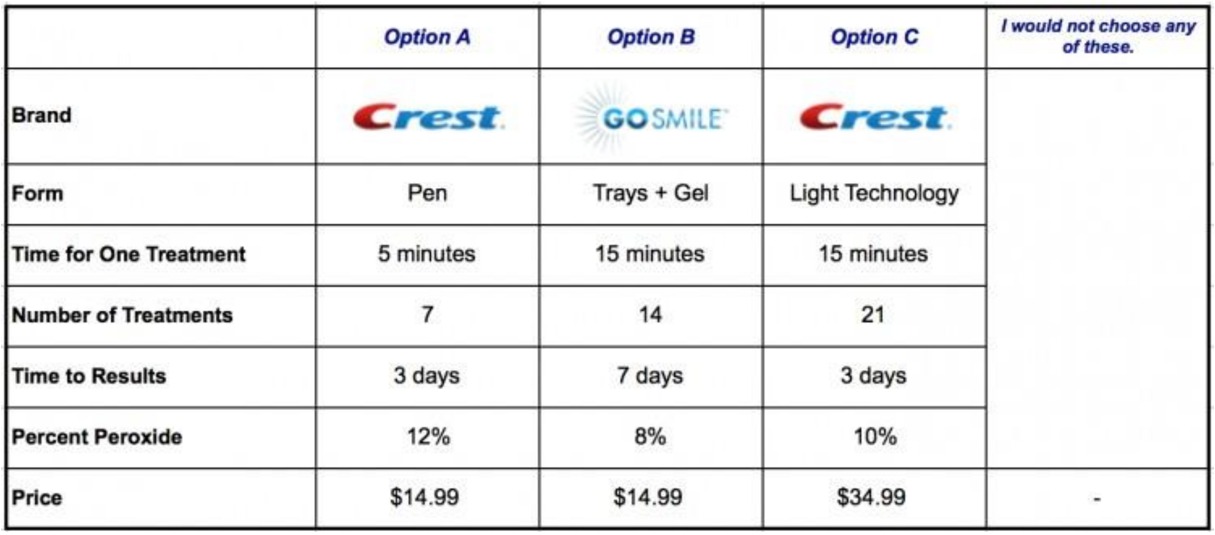
The third experimental condition provided information about the relationship between brand, price, and the effectiveness attributes. Graphs comparing price to time per treatment, number of treatments, time to results, and percentage of peroxide were explained to respondents and show in succession. Figure 2 displays the price versus time per treatment graphic shown to respondents.

Figure 2: Price vs. Time per Treatment Graphic



Conjoint choice tasks were then presented to the qualified respondents, asking them to identify their most preferred offering from among those chosen. Each choice task included three different brand names and an outside option from which respondents indicated their preferred brand. Each respondent was asked to indicate their preferences across 12 choice tasks. The attribute levels were experimentally manipulated across the choice tasks according to principle of statistical experimental design (Box et al., 1978) so that they were statistically identified. An example choice task is shown in Figure 3.

Figure 3: Example Choice Task



	Option A	*Option B*	*Option C*	*I would not choose any of these.*
Brand	Crest	GO SMILE	Crest	
Form	Pen	Trays + Gel	Light Technology	
Time for One Treatment	5 minutes	15 minutes	15 minutes	
Number of Treatments	7	14	21	
Time to Results	3 days	7 days	3 days	
Percent Peroxide	12%	8%	10%	
Price	$14.99	$14.99	$34.99	-

The survey ended with a series of demographic variables to help assess the representativeness of the sample population. A total of 1141 qualified respondents completed the survey and provided information about their preferences. The number of respondents was evenly split across the three experimental conditions.

3 Parameter Estimates

During estimation, two additional data screens were employed to remove respondents who did not appear to take the survey seriously. Respondents were screened from inclusion of the study if they were found to straight-line their responses by always selecting a choice alternative in the same position (e.g., the left-most) in the choice task. This eliminated 44 respondents from the analysis. An additional, model-based data screen was implemented to guard against respondent guessing. This was accomplished by fitting a hierarchical multinomial logit model to the full dataset and using the resulting individual log likelihood values to screen out guessing respondents. Guessing produces unreliable estimates of part-worths and price responsiveness that artificially inflates estimates of the economic value of the part-worths. Guessing can be detected by low probabilities of choice as predicted by the model. A person providing random answers to the choice task would lead to low predicted choice probabilities. By analyzing only those respondents who were somewhat predictable in their choices, one can obtain demand estimates and assess model results that reflect an engaged set of respondents.

The conjoint tasks in the survey involve making choices from among three alternative teeth-whitening products and a no-choice option. A person guessing at random among the four alternatives would result in a naïve choice probability of 0.25 (i.e., 1/4) for each of the alternatives. As a respondent's choices become more predictable, the probability of the chosen alternative increases. We set a minimum threshold of 0.40 as the average choice probability for a respondent. That is, respondents with an average fitted choice probability of 0.40 or less are screened out of the dataset, while those with an average choice probability above 0.40 are retained. This additional data screen resulted in 12 respondents being eliminated from analysis, leaving a total of 1085 respondents across the three experimental conditions.

In all three experimental conditions the “I would not choose any of these” option was chosen around 20% of the time, indicating similar preferences for the outside good across conditions. Table 2 reports the posterior mean and standard deviation of the random-effects distribution for the three experimental conditions— condition 1 where no information is provided about competitive prices, condition 2 where price information is provided, and condition 3 where price and attribute information is provided. Approximately 95% of the mass of heterogeneous responses is within plus or minus two posterior standard deviations of the mean. A comparison of the random-effects distribution for the three conditions indicates similarity in the outcome measures, implying that informing respondents of a broader array of marketplace prices and attribute performance has little effect on the estimated part-worths. A more in-depth study of differences due to the experimental conditions, as measured in economic terms, is presented below in Section 5.

4 Economic Measures of Value

We now compare the parameter estimates reported in Table 3 using various measures of economic value used to assess part-worths in a conjoint study (see Allenby et al., 2019; Lloyd-Smith, 2018). To begin this analysis we will start with measures of willingness-to-pay, a demand-based estimate of monetary value.

Table 3: Posterior Mean and Standard Deviation of Heterogeneity

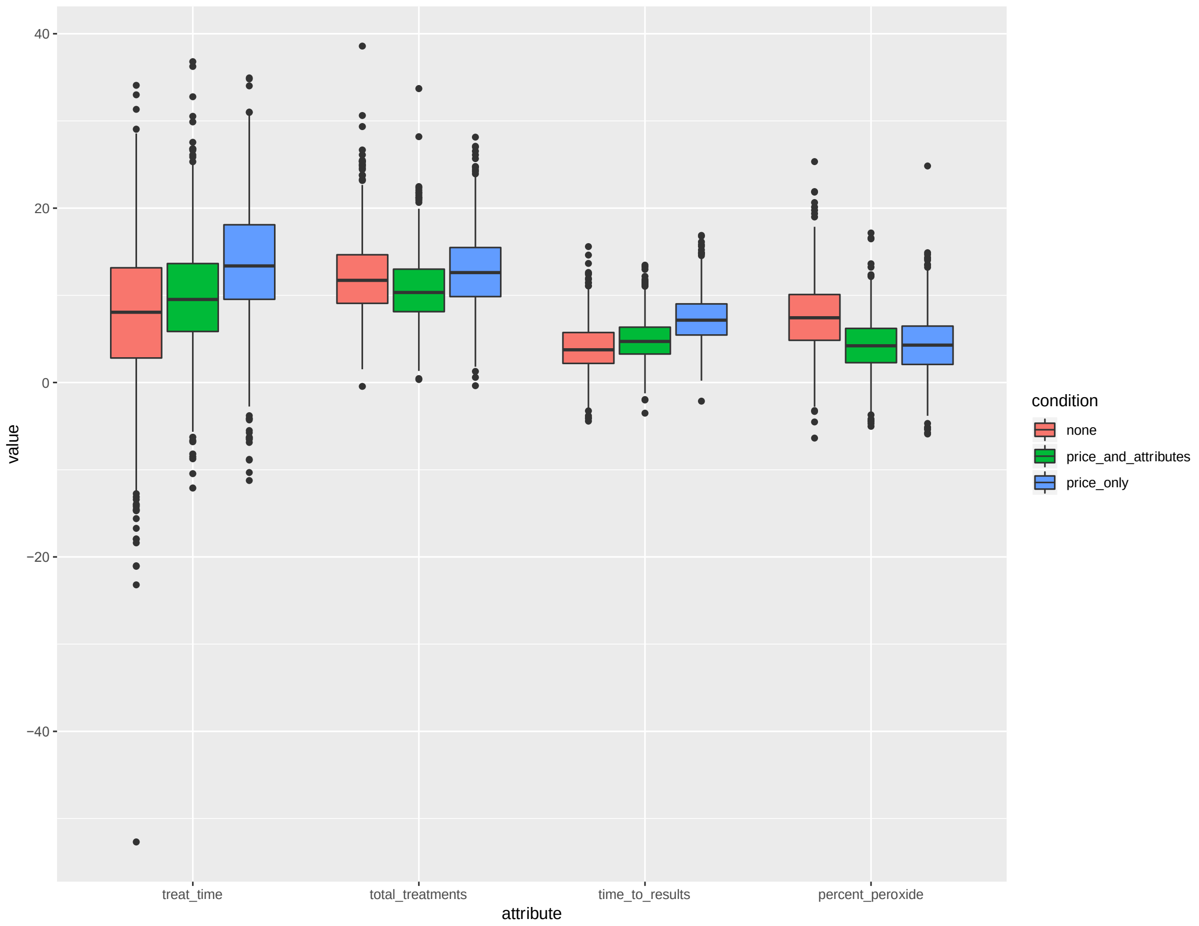
	No Information		Price Only		Price and Attributes	
Attribute Levels	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept	19.374	12.281	18.258	11.457	20.354	11.938
Plus White	−2.259	3.549	−1.364	4.145	−2.389	3.815
Rembrandt	−1.774	4.230	−1.631	4.412	−1.366	3.957
Go Smile	−1.472	2.875	−0.860	4.146	−0.833	3.908
Auraglow	−2.778	4.912	−2.196	4.548	−3.212	3.705
Luster	−2.356	3.911	−2.436	4.873	−2.747	3.714
Pen	−1.392	5.328	−1.185	5.223	−1.298	5.470
Trays & Gel	−1.052	4.752	−1.399	6.068	−1.366	5.159
Light Tech	−1.794	6.562	−1.678	6.280	−2.099	6.468
Treatment Time	−0.193	0.460	−0.194	0.430	−0.212	0.469
No. Treatments	−0.138	0.362	−0.147	0.349	−0.115	0.366
Time to Results	−0.085	0.412	−0.109	0.381	−0.141	0.411
% Peroxide	0.247	0.615	0.209	0.661	0.130	0.672
Price	−1.754	1.361	−1.904	1.241	−1.669	1.263
Sample Size	N=379		N=371		N=355	

4.1 Pseudo willingness to pay (p -WTP)

A naïve measure of WTP, which does not consider the set of competitive offers, is a simple monetization of utility to a dollar measure. We refer to this as a pseudo measure (p-WTP) because it assumes that consumers are locked into the purchase of a specific brand (e.g., Crest) and ignores the fact that consumers can switch brands and decide not to make a purchase. The measure is constructed using the ratio of attribute-level part-worth and the price coefficient, p-$WTP = \beta_i/\beta_p$. To be clear, we do not advocate the use of this measure, but include it for comparison purposes.

Figure 4 displays boxplots for the p-WTP measure of product improvement for the Crest brand. The boxes in the plot correspond to the inter-quartile range containing 50% of the distribution for each measure. The whiskers of the boxplot correspond to 1.5 times the inter-quartile range, and the extending points are outliers. The uncertainty displayed in the plots is due to uncertainty in the parameter estimates in the hierarchal Bayes model. The p-WTP measure varies between $5.00 and $10.00 for each of the attributes and is similar across treatment conditions. That is, the effect of information about outside goods does not appear to be large or consistent in its effect, especially when considering the uncertainty in the estimates.

Figure 4: p-WTP for Crest Product Improvement



4.2 Economic willingness to pay (WTP)

An alternative WTP measure that corresponds to a measure of consumer welfare acknowledges the presence of alternatives with non-zero choice probabilities by measuring the maximum attainable utility from a transaction. Increasing the number of available choice alternatives increases the expected maximum utility a consumer can derive from a marketplace transaction, and ignoring the effect of competitive products leads to a misstatement of consumer welfare. Any measurement of the economic value of a product feature cannot be made in isolation of the set of available alternatives because it is not known, a priori, which product will be chosen.

Measurement of the economic value of a product attribute requires the specification of a set of products for which choices are predicted. This is accomplished by constructing two choice sets. The first choice set is described by the matrix A with rows describing choice alternatives and columns detailing which of the alternatives has each product feature under study. The choice matrix A^* is similar to A except that one of its rows is different, indicating a different set of features for one of the alternatives. Typically, just one element in the row differs when comparing A to A^* because the economic measure typically focuses on what respondents are willing to pay for an enhanced version of one of the attributes.

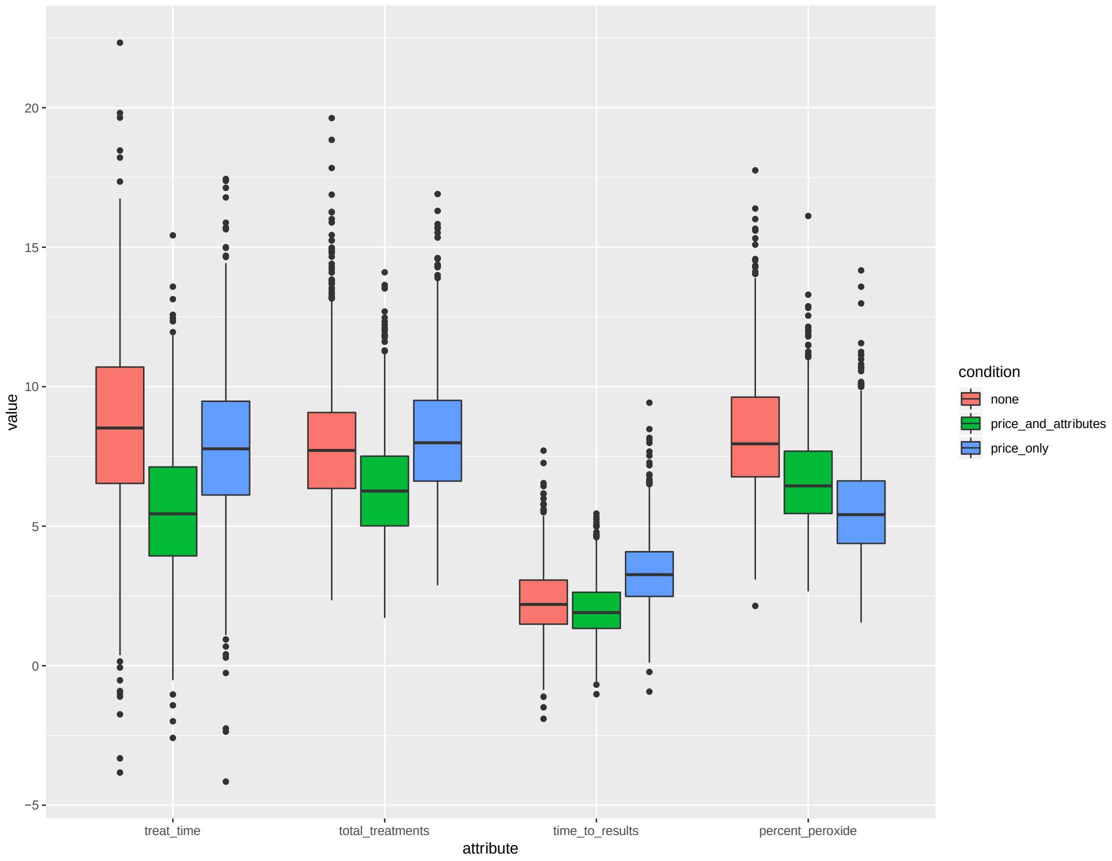
We consider scenarios where the competitive set is comprised of 6 brands (Crest, Plus White, Rembrandt, GoSmile, Auraglow, and Luster), with each offered using whitening strips technology and having an average level of performance for each attribute; i.e., treatment times of 15 minutes, the suggested number of treatments is 14, results in 7 days, 10% peroxide content, at a price of $34.99. The matrix A^* is identical to A except the row for Crest is altered, one attribute at a time, to value an improved level of performance: treatment times of 5 minutes, the suggested number of treatments is 7, time to results is 3 days, and 13% peroxide content.

As discussed by (Small and Rosen, 1981; Lancsar and Savage, 2004; Allenby et al. 2014a), we construct the maximum level of utility attainable for the improved choice set A^* and the original choice A and calculate the difference.[2] Figure 5 displays boxplots for the WTP that accounts for uncertainty in choice and the presence of competitive products. The measures of economic value are smaller than the p-WTP measures, as expected, and there is greater distinction in value across the attributes. However, the distinction among the experimental conditions remains small and inconsistent; e.g., sometimes the None experimental condition leads to highest economic value and other times it does not, especially given the uncertainty in the estimates.

[2] $\text{WTP} = \frac{\ln\left[\sum_{j=1}^{J} \exp\left(\beta_h' \boldsymbol{a}_j^* - \beta_{hp}' \boldsymbol{p}_j\right)\right]}{\beta_{hp}} - \frac{\ln\left[\sum_{j=1}^{J} \exp\left(\beta_h' \boldsymbol{a}_j - \beta_{hp}' \boldsymbol{p}_j\right)\right]}{\beta_{hp}}$

where a_j indicates the features of the j^{th} product in the corresponding choice set matrix.

Figure 5: WTP for Crest Product Improvement



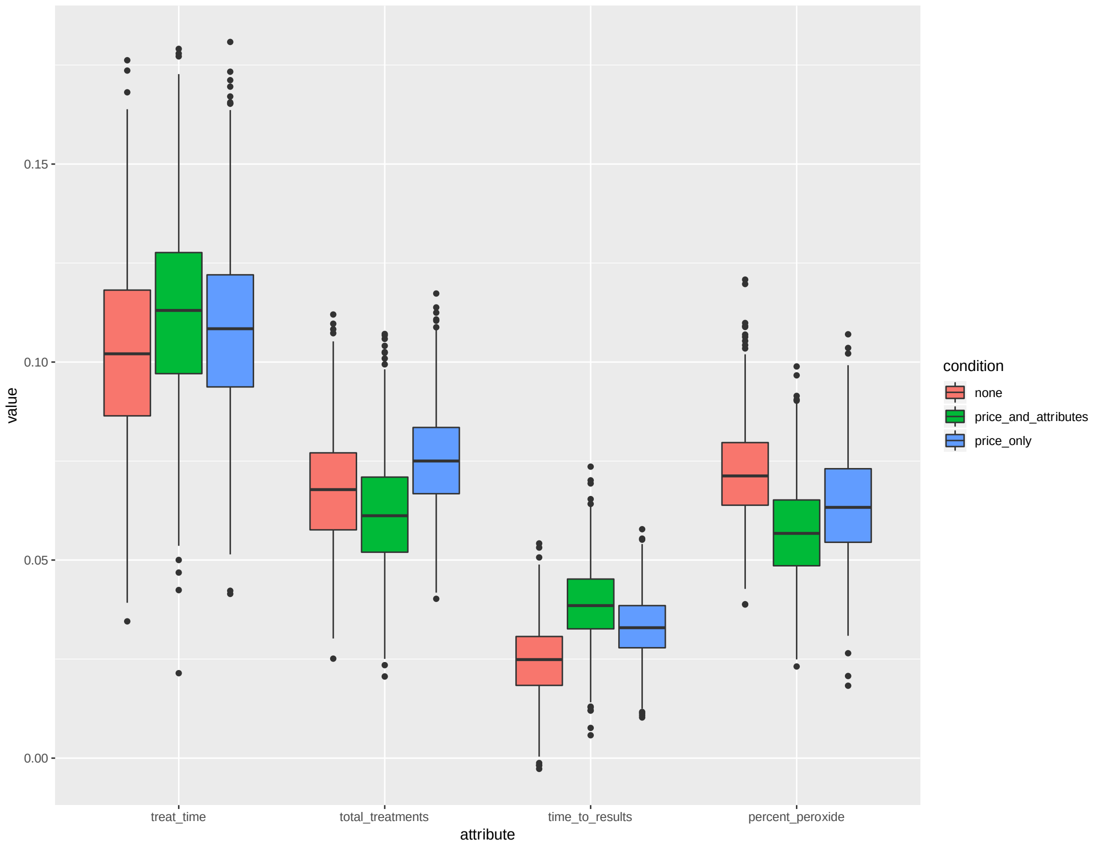
4.3 Willingness to Buy (WTB)

Willingness to buy is a measure of economic value based on the expected increase in demand for an enhanced offering. Economic value is determined by calculating the expected increase in revenue or profit due to a feature enhancement, using WTB as an input to that calculation. The increase in demand due to the improved feature is calculated for one offering, holding fixed all of the other offerings in the market. In discrete choice models, WTB is defined in terms of the change in share that can be achieved by moving from the original feature set to an improved feature set for the product.[3]

Figure 6 displays boxplots for the improvements to Crest whitening strips in terms of aggregate choice shares. The increase in share is largest for the treatment time attribute, where an improvement from 15 minutes to 5 minutes is estimated to an improvement of 0.10. The increase is smallest for the time-to-results attribute, where an improvement from 7 days to 3 days leads to an increase in share of about 0.03. The posterior distribution of the share increase overlaps greatly for each attribute improvement, indicating that the effect of outside good information in the form of competitive prices in Figure 1 and/or attributes as in Figure 2 is small.

[3] $\text{WTB} = \text{MS}(j|\mathbf{p},\mathbf{A}^{*}) - \text{MS}(j|\mathbf{p},\mathbf{A})$s

Figure 6: WTB for Crest Product Improvement



4.4 Economic Price Premium (EPP)

The economic price premium is a measure of feature value that allows for competitive price reaction to a feature enhancement. There are two advantages for allowing competitive prices to adjust to a feature improvement. First, the resulting equilibrium calculation measures the long-term effect of a feature enhancement after the market has adjusted to the improvement. Second, the EPP estimates tend to be smaller and more realistic that the WTP estimates that do not allow for competitive reactions. The effect of a competitive reaction is always to decrease, not increase, the economic value of a product improvement to a firm.

An equilibrium is defined as a set of prices and accompanying market shares which satisfy the conditions specified by a particular concept of equilibrium. We employ a Nash Equilibrium concept for differentiated products using a discrete choice model of demand. The calculation of an equilibrium price premium requires additional assumptions beyond those employed in a traditional conjoint study:

- The demand specification is a standard heterogeneous logit that is linear in the attributes, including prices.
- Constant marginal cost for the product.
- Single product firms; i.e., each firm has just one offering.

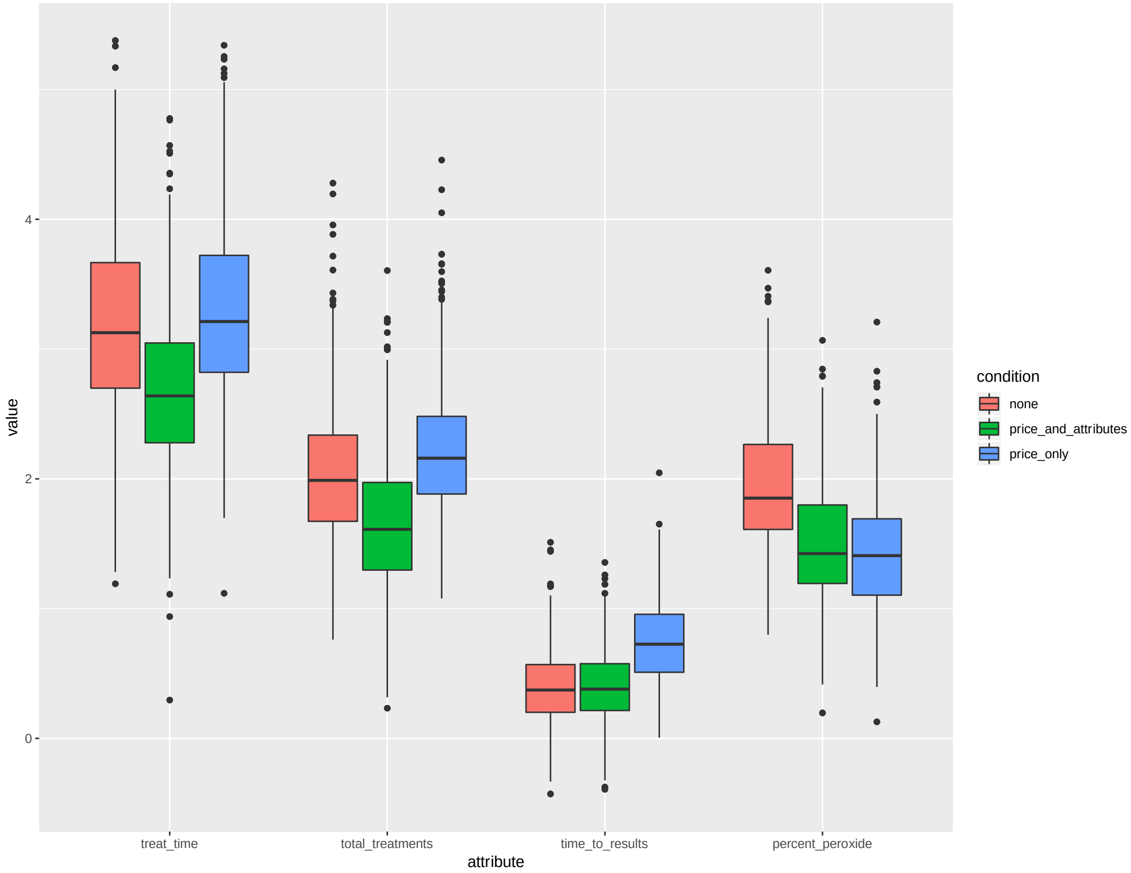
- Firms cannot enter or exit the market after product enhancement takes place.
- Firms engage in a static Nash price competition.

The economic value of a product feature enhancement is the incremental profits that it will generate. That is, the change in profits, π, associated with the equilibrium prices and shares given a set of competing products defined by the attribute matrix A. This can be constructed by finding the first-order conditions with respect to the profit function for each firm.[4] The Nash equilibrium is a root of the system of equations defined by these first-order conditions for all J firms.

Figure 7 displays boxplots for the improvements to Crest in terms of the economic price premium. As expected, the value attributed to the feature enhancements is estimated to be less than the p-WTP measure displayed in Figure 4 and the WTP measure displayed in Figure 5. The EPP value for an improvement in treatment time is about \$3.00 rather than \$8.50, and the value for an improvement in time-to-results is estimated to \$.40 rather than \$2.20. However, consistent with the other measures of economic value, the effect of the different treatment conditions is small.

[4] $\frac{\partial \pi}{\partial p_j} = \frac{\partial}{\partial p_j} MS\left(j \middle| p_j, \boldsymbol{p}_{-j}, \boldsymbol{A}\right)\left(p_j - c_j\right) + MS\left(j \middle| p_j, \boldsymbol{p}_{-j}, \boldsymbol{A}\right)$
where pj is the price of good j, p−j are the prices of other goods, and cj is the marginal cost of good j.

Figure 7: EPP for Crest Product Improvement



5 Conclusion

The preliminary results above suggest that conjoint part-worth estimates are robust to the amount of information about marketplace prices and feature performance that is provided to respondents. We find no systematic difference in parameter estimates as reported in Table 2 nor in economic measures of value as displayed in Figures 4 to 7. Moreover, posterior estimates of the effects, as displayed in the figures, show a great deal of overlap across treatment conditions. Since our study screens respondents for inclusion if they are active in the product category this may come as no surprise.

Our analysis lends support to the economic view of choice in that qualified respondents were able to recall and construct their preferences for the choice alternatives without being influenced by the price and attribute information provided in the glossary video. We did not detect effects due to priming (Shrum et al., 1998) and other mechanisms where respondents exposed to attribute information are induced to make choices favoring the certain attributes. Respondents were able to recall and form their preferences for the conjoint offerings independent from the information provided in the survey. We believe that the reason for the robustness of results is due to the respondents being familiar with the product category and product attributes prior to engaging in the study.

Given the difficulty in "proving" such a result, we plan to continue the project by fleshing out the full experimental design in additional product categories, and investigating the outcomes for respondents that would not have passed the study screens. We also plan to investigate the amount of marketplace information that is internalized by respondents to see if the presentation of this information results in increased respondent familiarity with the current state of the category. Additional research also is needed to understand boundary conditions for obtaining stable estimates in conjoint studies and surveys in general, especially in relation to best practices regarding screening procedures.

Maggie Chwalek

Roger A. Bailey

Greg M. Allenby

References

Allenby, Greg M, Jeff Brazell, John R Howell, Peter E Rossi. 2014a. Valuation of patented product features. *The Journal of Law and Economics* **57**(3) 629–663.

Allenby, Greg M, Jeff D Brazell, John R Howell, Peter E Rossi. 2014b. Economic valuation of product features. *Quantitative Marketing and Economics* **12**(4) 421–456.

Allenby, Greg M, Nino Hardt, Peter E. Rossi. 2019. *Handbook of the Economics of Marketing,* chap. Economic Foundations of Conjoint Analysis. Elsevier.

Bahamonde-Birke, Francisco J., Isidora Navarro, Juan de Dios Ortúzar. 2017. If you choose not to decide, you still have made a choice. *Journal of Choice Modelling* **22** 13–23.

Balbontin, Camila, David A. Hensher, Andrew T. Collins. 2017. Do familiarity and awareness influence voting intention: The case of road pricing reform? *Journal of Choice Modelling* **25** 11–27.

Ben-Akiva, Moshe, Daniel McFadden, Kenneth Train. 2018. Foundations of stated preference elicitation. Working paper.

Box, George EP, William Gordon Hunter, J Stuart Hunter, et al. 1978. *Statistics for experimenters*. John Wiley and sons New York.

Brazell, Jeff D, Christopher G Diener, Ekaterina Karniouchina, William L Moore, Válerie Séverin, Pierre-Francois Uldry. 2006. The no-choice option and dual response choice designs. *Marketing Letters* **17**(4) 255–268.

Carson, Richard T., Jordan J. Louviere, Donald A. Anderson, Phipps Arabie, David S. Bunch, David A. Hensher, Richard M. Johnson, Warren F. Kuhfeld, Dan Steinberg, Joffre Swait, Harry Timmermans, James B. Wiley. 1994. Experimental analysis of choice. *Marketing Letters* **5**(4) 351–367.

Green, Paul E, Vithala R Rao. 1971. Conjoint measurement for quantifying judgmental data. *Journal of Marketing research* 355–363.

Lancsar, Emily, Elizabeth Savage. 2004. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. *Health economics* **13**(9) 901–907.

Lichtenstein, Sarah, Paul Slovic. 2006. *The construction of preference*. Cambridge University Press.

Lloyd-Smith, Patrick. 2018. A new approach to calculating welfare measures in kuhn-tucker demand models. *Journal of choice modelling* **26** 19–27.

Louviere, Jordan J, Terry N Flynn, Richard T Carson. 2010. Discrete choice experiments are not conjoint analysis. *Journal of Choice Modelling* **3**(3) 57–72.

Manski, Charles F, Daniel McFadden, et al. 1981. *Structural analysis of discrete data with econometric applications*. Mit Press Cambridge, MA.

Rao, Vithala R. 2014. *Applied conjoint analysis*. Springer.

Sandorf, Erlend Dancke, Danny Campbell, Nick Hanley. 2017. Disentangling the influence of knowledge on attribute non-attendance. *Journal of Choice Modelling* **24** 36–50.

Shrum, LJ, Robert S Wyer Jr, Thomas C O'Guinn. 1998. The effects of television consumption on social perceptions: The use of priming procedures to investigate psychological processes. *Journal of Consumer Research* **24**(4) 447–458.

Small, Kenneth A, Harvey S Rosen. 1981. Applied welfare economics with discrete choice models. *Econometrica: Journal of the Econometric Society.*

MODELLING STOCKPILABLE PRODUCT PURCHASE DECISIONS USING VOLUMETRIC CHOICE EXPERIMENTS[1]

RICHARD T. CARSON
UNIVERSITY OF CALIFORNIA, SAN DIEGO[2]
TOWHIDUL ISLAM
UNIVERSITY OF GUELPH, CANADA[3]
JORDAN J. LOUVIERE
UNIVERSITY OF SOUTH AUSTRALIA[4]

ABSTRACT

Individuals often decide how many units of a specific good to purchase rather than simply deciding whether to purchase it or not. There is much more information in such count data than in traditional discrete choices. We consider a shelf of canned tuna and survey a sample taken from the Information Resources, Inc. (IRI) consumer panel who met the condition that they had purchased at least one can of tuna in the prior year. We designed and implemented a volumetric choice experiment (VCE) to obtain 120,000 quantity choices (defined by price, brand, size, and other attributes) in a stated choice context. Canned tuna is the quintessential stockpilable commodity so it may be particularly sensitive to price and promotions like savings coupons for specific brands. The VCE experimental design we used allows for statistical identification of over 100 own (brand by size) and cross price elasticities. We discuss preliminary estimates from a multilevel mixed-effects negative binomial regression.

INTRODUCTION

This paper examines choice opportunities, where instead of facing a single discrete choice between a set of competing alternatives, consumers face a choice of how many units to purchase, if any of several alternatives that are available. The quintessential example that we address here is different types of canned tuna on a grocery store shelf. While the two situations have key similarity including being amenable to the use of experimental design to identify key demand parameters, they are substantively different. The ability to observe discrete volumetric choices allows the estimation of a much richer set of consumer demand models.

Economic theorists initially conceived of consumer demand in terms of continuous quantities for two reasons: 1) the underlying mathematics is much easier than alternatives that allow for various types of discreteness; and 2) the only available data for empirical work at the time was aggregate in nature, so the discrete

[1] We gratefully acknowledge research support from Social Sciences and Humanities Research Council (SSHRC) of Canada, Grant No. 430199.

[2] Professor of Economics, University of California, San Diego; Chief Scientist, ChoiceFlows, Research Triangle Region, NC

[3] Professor of Marketing and Consumer Studies, University of Guelph, Canada

[4] Emeritus Research Professor of Marketing, University of South Australia; Chief Operating Officer & Co-Founder, ChoiceFlows, Research Triangle Region, NC

aspects of demand did not seem to matter. Constraints that theoretically should apply only to an individual's demand, such as adding up restrictions, were routinely imposed on demand systems estimated using the available aggregate data (Phlips, 1974).

When individual-level data became available, attention naturally turned to the modelling of discrete choices, with the random utility model (RUM) becoming the workhorse of empirical analysis (McFadden, 1974). Over time, those working with discrete choice data moved to address two key limitations of the aggregate conditional logit model. The first is the restrictive assumption that all agents have the same preferences except for an idiosyncratic error term that followed a standard Type I extreme value distribution. These advances involved new functional forms that relax the conditional logit's independence of irrelevant alternatives (IIA) assumptions and provide explicit ways to allow for preference heterogeneity in both a frequentist (Train, 2009) and Bayesian context (Rossi, Allenby and McCulloch, 2005). The second involves ways to address the limitations of revealed preference purchase data by proposing econometric approaches to deal with endogeneity (Berry, Levinsohn and Pakes, 1995), particularly with respect to product prices. It also involves using stated preference data along with an experimental component that solves the endogeneity problem by randomly assigning product attribute levels including price using what are known as discrete choice experiments (DCEs) (Louviere, Hensher and Swait, 2000). This paper contributes to the third wave in modelling consumer demand, where that demand is both continuous and discrete in the sense that demand takes the form of the set of non-negative integers; i.e., the dependent variable is a volumetric choice that statistically can be represented as a count data regression model (Cameron and Trivedi, 2013).

Many examples of volumetric choices easily come to mind, such as the number of frozen yogurts bought during a trip to the supermarket or the number of polo shirts purchased from an online retailer. Other examples include the number of credit cards in someone's wallet or the number of social media accounts that they have; the number of stores visited during a trip to a mall or the number of distinct machines played during an arcade stop. So, many discrete activities become count data when observed over a time interval, such as the number of flights taken last month or the number of primary care visits last week. There is a long history (e.g., Goodhardt and Ehrenberg, 1967; Schmittlein, Bemmaor and Morrison, 1985) of using volumetric choice models for revealed preference data in applied economic work including marketing, but the literature is dramatically smaller than that involving discrete choice models. More limited still is work on volumetric choice using stated preference data. Early examples (Carson, Hanemann and Steinberg, 1990) tended to offer survey respondents choices that represented the discrete number of a good they could purchase. These data then were modelled using a range of discrete choice models that effectively ignored most, if not all, of the statistical properties of count data. Now one can design volumetric choice experiments (VCEs) (Louviere, Ribeiro and Carson, 2016) that are cousins of the popular DCEs (Louviere, Hensher and Swait, 2000). In this paper we discuss the design of and preliminary estimates from a large scale VCE that examined how consumers choose when faced with a grocery store shelf of different types of canned tuna, a textbook example of a stockpilable commodity.

Volumetric Choice: Econometric Models

Count data models can be used to fit the decisions made in VCEs. Cameron and Trivedi (2013) provide a comprehensive overview of count data models used in applied economic work.[5] Two count data models are in general use. The first is the Poisson, whose key property in a regression context is that the conditional mean is constrained to be equal to the conditional variance. Empirically, this restriction rarely holds with consumer demand data due to over-dispersion (conditional variance > conditional mean) being typical. If one is solely interested in the parameters of the demand function, the mean-variance constraint is not as restrictive as it might first appear because the Poisson is still a quasi-maximum likelihood estimator if the conditional mean function is correctly specified, and consistent standard errors can be obtained by using White-type standard errors.

The main alternative to the Poisson is the negative binomial model, which explicitly parameterizes the variance in various ways that differ in terms of flexibility and computational tractability. A major attraction of the negative binomial model is the potential ability to better model the variability of the overall distribution of purchase counts and predictions for a given set of covariates. While the Poisson separates preference and scale parameters, that only appear in ratio form in discrete choice models, it does so with a very strong assumption. Variants of the negative binomial specification relax that constraint in different ways. Both the Poisson and negative binomial models are linear-in-the-parameters models, an assumption that is relaxed in their generalized additive model formulations. Importantly, for all count data models rather than the estimated coefficients being the ratio of preference parameters to the standard error as is the case for discrete choice models, count data models provide separate estimates for both preference parameters and the scale parameter.

The way(s) in which volumetric data are used in applied economics often results in the need for modification to the likelihood functions being maximized. Data is right censored if one only knows at the high end is that the number of goods purchased is larger than some quantity. This can occur in SP data when the last response category is specified as "X or more units," where X is some integer (e.g., 8). Left truncation at zero is common when only information on purchasers is

[5] As counts become large, the standard linear regression model tends to provide an adequate fit, so the formal use of count data models is typically appropriate in situations where small counts are common. Implicit in all the count data models considered here is that all goods are well defined in terms of their attributes and the counts represent ratio scale data. When used to model choices involving counts, the major issues with other commonly used alternatives are as follows. By construction, a conditional logit model of count data violates its underlying IIA assumption; the ordered logit/probit, which potentially produces consistent estimates, is inefficient because it has to estimate a set of threshold parameters; the nested logit model effectively deals with a zero-inflated component but its underlying conditional logit specification is violated with count data; ordinal regression foregoes the advantage that count data is ratio scaled; censored regression models like the Tobit allow for the possibility that true quantities purchased are negative, which is not possible; discrete-continuous models are similar to the nested logit in potentially dealing correctly with zero-inflated situations, but suffer the same problem as OLS with small counts. It is important to note that count data models assume that the attributes of goods including those like package size are fixed. When the purpose of the statistical analysis is to provide guidance on how to package or size goods because only a limited number of variants can be offered, a different statistical approach should be used. See Lee and Allenby (2014) an instructive application.

available so that zeros are not observed. When zeros are observed, so-called zero-inflated Poisson or negative binomial models are often estimated that allow the zeros to be generated by the same process as that generating the positive counts (statistical zeros) or structural zeros, which occur when no change in a good's attributes such as price is capable of moving the zero to a positive integer. The zero-inflated models are special cases of hurdle count data models which can be useful when there is interest in the process that leads to the eventual generation of positive counts. As multiple volumetric choice opportunities are observed, models can be estimated that allow for individual-level heterogeneity using individual-level fixed effects, random parameters specification for one or more of the model's covariates, or some variant of a latent class representation.

It is possible to allow for correctable endogeneity of a covariate in count data models if a suitable instrument is available, although randomization of attribute levels in a VCE obviates that problem. One also can allow for different types of correlation, including choices by the same agent over time and across choices involving similar goods at the same time. Count data models are available in both frequentist and Bayesian paradigms. A wide-range of standard and modified (e.g., censored, zero-inflated) count data models are available in R and Stata, as well as many other statistics packages.

Hellerstein and Mendelsohn (1993) show a simple Poisson demand model can be obtained in two distinct ways. The first is as a simple linear model of continuous quantity demand with the imposition of non-negative integer quantities constraints. The second is as a repeated binary choice model with any of the standard discrete choice models that assume errors are i.i.d. across choice occasions at an individual level. A straightforward example is a trip to Walmart taken or not on each day of a specific week, with the volumetric choice being the number of trips taken to Walmart over the course of that week. However, this also is consistent with a number of very common but less likely obvious situations that arise from asking a question like "How many times did the individual reach for a gallon of milk and put it into a shopping cart?," where the volumetric choice is the total number of gallons of milk.

Poisson regression models are obtained by letting the Poisson mean-variance parameter λ be a function of a set of observed covariates (Greene, 2003; Cameron and Trevedi, 2013). Specifically,

$$Prob(X_i) = \frac{e^{\lambda_i}\lambda_i^{y_i}}{y_i!}, y_i = 0,1,2 \text{ .... and i=1,2,3…. n,} \qquad (1)$$

The $y_i = 0,1,2$ .... are the realized values of the random variable, λ_i is the mean and variance of y_i., and X_i is a covariate vector. The most common formulation of the mean vector is $\lambda_i = e^{X_i'\beta}$. An illustrative example consistent with an underlying linear logistic model for each purchase/not purchase decision is:

$$\lambda = \exp[\beta_0 + \beta_1\text{Price} + \beta_2\text{Income} + \beta_3\text{Attributes} + \beta_4\text{Demographics}]. \quad (2)$$

Other i.i.d. choice models have Poisson demand, but λ may have a somewhat different functional form; and other functional form assumptions such as the conditional mean being linear in log price rather than price have well-known mapping in count data models. Negative binomial models use the same conditional mean specifications as the Poisson but allow for a separate scale parameter which effectively looks like tacking on a (gamma distributed) error term to some transformation of (1). This scale parameter also can be specified as a function of covariates, but this typically is not done due to the computational complexity introduced. The negative binomial model also avoids the independence assumption of the Poisson model, which has parallels to the conditional logit model's IIA.

With no income effect, the ordinary Marshallian consumer surplus estimate for the Poisson model in (1) is $-\lambda/\beta_1$. Marshallian and Hicksian welfare measures (e.g., maximum willingness to pay) for more complex count data models can be derived in a manner similar to that used for discrete choice models (Carson and Hanemann, 2005).

Product Category, Attributes, VCE Design, and Sample

We focus on canned tuna, long a staple of American shopping baskets, and for that reason it's often used to illustrate new marketing research approaches (e.g., Allenby, 1990). It is an easily storable commodity, which makes it ideal for looking at whether changes in price and the presence of promotions like coupons lead to substantial changes in the quantity purchases. Other features of canned tuna, which make it ideal for our purposes is that the three major brands provide over 90% of store purchase volumes (Starkist, Bumble Bee, and Chicken of the Sea) (store brands such as Walmart and Kroger capture much of the rest). This allows us to have a VCE with five effective brands: Starkist, Bumble Bee, Chicken of the Sea, store brand, and "Any other Brand" (e.g., Tonno Genova, Van Camp, Wild Planet). Canned tuna is effectively characterized by two sizes, small and large, being packed in oil versus water, being Albacore or another type of tuna, and form (chunk, i.e, flakes or solid).

Each brand option is represented by two sizes (large and small), and each brand had five attributes (price, coupon, type of tuna, packaging and form), where the last three attributes are binary. Each brand and size were represented by a 2^3 x 8^2, so the full factorial is a (2^3 x 8^2)^10, or a 2^30 x 8^20. An additional 16-level column was used to create blocks (versions). An orthogonal main effects design in 256 rows was used to make the choice sets. We used an Alternative-Specific Design (Louviere and Woodworth, 1983) because it forces the attributes of each alternative that can be chosen to be orthogonal both within and between alternatives. This feature insures "own effects" (e.g., own elasticities) within each alternative and "cross effects" (e.g., cross-elasticities) between alternatives are statistically identified. Commonly used generic DCEs or VCEs generally do not have this property.

Although canned tuna is available in cans of different sizes, the vast majority of sales are in 4 to 6 oz. cans and 10 to12 oz. cans. So, we offered 6 and 12 oz. sizes to standardize the small/large dichotomy in the market. This creates 10 options in each choice set. We also varied price (combined with size as one attribute), what the tuna was packed in (oil or water), the type of tuna (Albacore or "Tuna") and the form of the tuna (Solid or Chunk). Finally, we also varied whether there was a coupon available, and if available the value it has. Price and coupon are 8-level attributes, all others are 2-level attributes. We also used a 16-level orthogonal blocking column to create blocks. Taken together, this produces a 16 x 8^10 x 2^15 factorial, and we selected the smallest orthogonal main effects design from that factorial, which produces 256 choice sets. As earlier noted, panelists were randomly assigned to one of the 16 blocks and then received the 16 choice sets in that block in a random order.

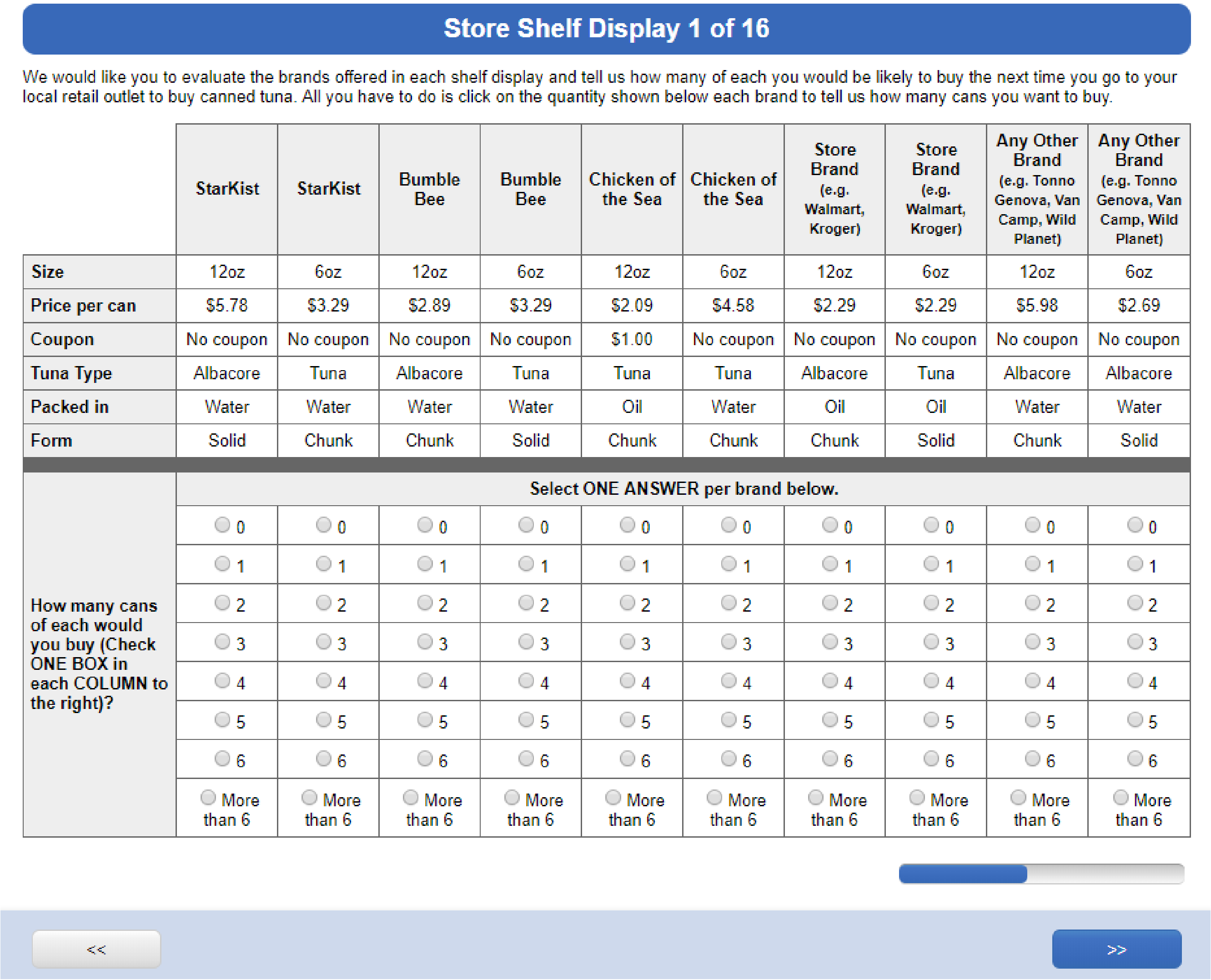
Table 1 shows the attributes and levels used in our VCE design for canned tuna. These attributes are assembled according to the design described above to form choices sets. Figure 1 displays a representative choice set faced by our respondents where our objective was to simulate a choice task like what an IRI panelist would see on a grocery store shelf of canned tuna. Our respondents were asked to indicate the number of cans they would be most likely to purchase in each of the 16 choice sets.

Table 1: Canned Tuna (Brand and Size: 10 alternatives):
Attributes and Their Levels

Attributes\Levels		Size (Oz)	Level 0	Level 1	Level 2	Level 3	Level 4
Brand/Price ($)	Starkist	6	2.69	2.89	3.09	3.29	
		12	5.38	5.78	6.18	6.58	
	Bumble Bee	6	2.89	3.09	3.29	3.49	
		12	5.78	6.18	6.58	6.98	
	Chicken of the Sea	6	1.99	2.09	2.29	2.39	
		12	3.98	4.18	4.58	4.78	
	Store Brands	6	1.89	2.09	2.29	2.49	
		12	3.78	4.18	4.58	4.98	
	Any Other Brands	6	2.69	2.99	3.29	3.59	
		12	5.38	5.98	6.58	7.18	
Type			Albacore	Tuna			
Packed in			Oil	Water			
Form			Chunk	Solid			
Coupon ($)**			0	0.50	0.75	1.00	1.50

** Proportions of levels different, about 77.5% $0 following revealed purchase data.

Figure 1: An Example of One Choice Set from the Canned Tuna VCE



Store Shelf Display 1 of 16

We would like you to evaluate the brands offered in each shelf display and tell us how many of each you would be likely to buy the next time you go to your local retail outlet to buy canned tuna. All you have to do is click on the quantity shown below each brand to tell us how many cans you want to buy.

	StarKist	StarKist	Bumble Bee	Bumble Bee	Chicken of the Sea	Chicken of the Sea	Store Brand (e.g. Walmart, Kroger)	Store Brand (e.g. Walmart, Kroger)	Any Other Brand (e.g. Tonno Genova, Van Camp, Wild Planet)	Any Other Brand (e.g. Tonno Genova, Van Camp, Wild Planet)
Size	12oz	6oz	12oz	6oz	12oz	6oz	12oz	6oz	12oz	6oz
Price per can	$5.78	$3.29	$2.89	$3.29	$2.09	$4.58	$2.29	$2.29	$5.98	$2.69
Coupon	No coupon	No coupon	No coupon	No coupon	$1.00	No coupon	No coupon	No coupon	No coupon	No coupon
Tuna Type	Albacore	Tuna	Albacore	Tuna	Tuna	Tuna	Albacore	Tuna	Albacore	Albacore
Packed in	Water	Water	Water	Water	Oil	Water	Oil	Oil	Water	Water
Form	Solid	Chunk	Chunk	Solid	Chunk	Chunk	Chunk	Solid	Chunk	Solid
	Select ONE ANSWER per brand below.									
How many cans of each would you buy (Check ONE BOX in each COLUMN to the right)?	○ 0	○ 0	○ 0	○ 0	○ 0	○ 0	○ 0	○ 0	○ 0	○ 0
	○ 1	○ 1	○ 1	○ 1	○ 1	○ 1	○ 1	○ 1	○ 1	○ 1
	○ 2	○ 2	○ 2	○ 2	○ 2	○ 2	○ 2	○ 2	○ 2	○ 2
	○ 3	○ 3	○ 3	○ 3	○ 3	○ 3	○ 3	○ 3	○ 3	○ 3
	○ 4	○ 4	○ 4	○ 4	○ 4	○ 4	○ 4	○ 4	○ 4	○ 4
	○ 5	○ 5	○ 5	○ 5	○ 5	○ 5	○ 5	○ 5	○ 5	○ 5
	○ 6	○ 6	○ 6	○ 6	○ 6	○ 6	○ 6	○ 6	○ 6	○ 6
	○ More than 6	○ More than 6	○ More than 6	○ More than 6	○ More than 6	○ More than 6	○ More than 6	○ More than 6	○ More than 6	○ More than 6

Respondents were a random sample of IRI's consumer panel who met the selection criteria of having purchased canned tuna at least once in the last year.[6] An important aspect of this selection criteria is that it eliminates the need to consider count data models that allow zero purchase observations to come from two distinct processes. Having purchased canned tuna in the past year rules out the possibility that observed zeros are structural rather than statistical in nature.[7] After dropping a small number of respondents who failed screening criteria (e.g., only purchased once in the first choice set [quitters] or failing to provide responses to all choice sets), our

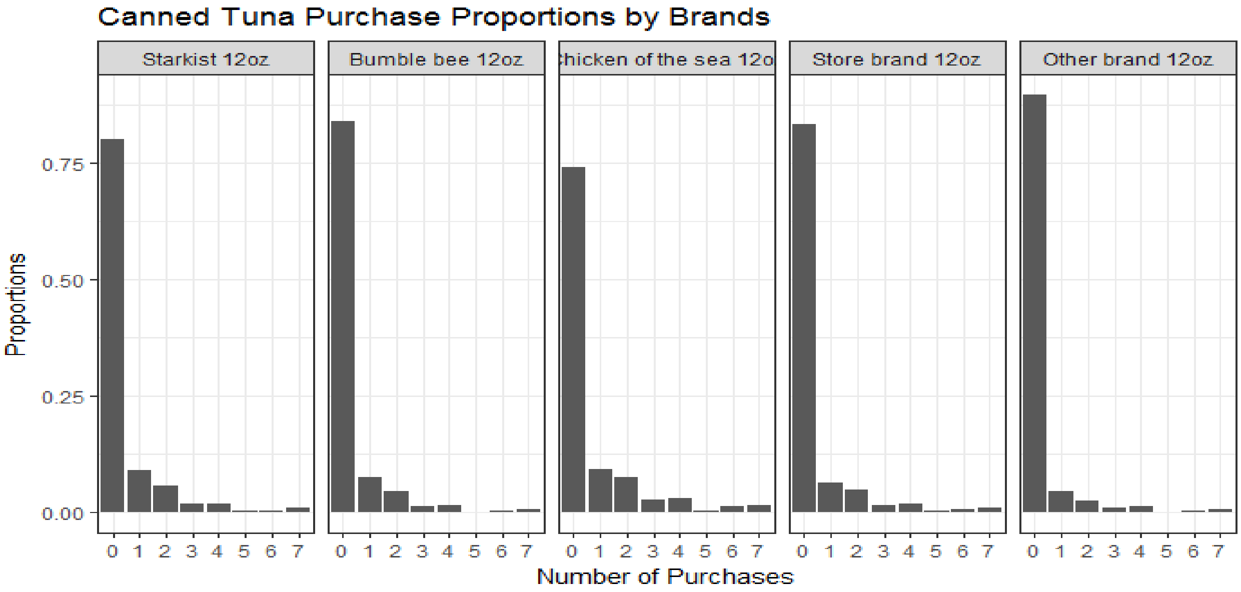
[6] We do not draw any inference here beyond our sample. However, the IRI Panel, with appropriate weights and accounting for stratification and clustering is intended to be representative of U.S. consumers. Since we have randomly sampled from it subject to a restriction (purchased canned tuna at least once in the last year), our results could be made representative of that restricted population of U.S. consumers, after rolling up the original IRI sample correction procedures with any subsequent deviations from simple random sampling that occurred with our sample.

[7] The canned tuna market is characterized by frequent specials involving sales and coupons that make it unlikely that a large drop in price will induce a substantial number of consumers, who did not purchase in the last year, to enter the market. This assumption may not hold for other products and, in general, modeling those who would potentially enter the market under the right conditions may be the most important part of a prospective demand analysis.

sample had 750 respondents. They were presented with 16 choice sets, each of which contained ten alternatives where the respondent picked a quantity, including zero, they would purchase. This produced 120,000 (750 x 16 x 10) volumetric choices.

Figure 2 displays the proportions of each count for the large (12 oz.) cans by brand in our VCE. Some of the key features are readily apparent. In most choice opportunities, respondents indicate that they would not purchase. Positive counts tend to decline monotonically as the count increases. A preference for Chicken of the Sea over other brands is apparent from its shorter histogram bar at zero. Results for the small sized cans are similar except for the overall propensity of the zero counts to be smaller and the positive counts higher, which is reflect of the strong preference for purchasing small versus large cans.

Figure 2



Comparison with Paired RP Data

Our sample is drawn from IRI's consumer panel and we obtained information on actual canned tuna purchases made by our respondents for the three-month period before our VCE was implemented and for three months afterwards. Potentially, this allows for a direct comparison between the SP data from our VCE and the RP data from the IRI consumer panel that would allow external validity of our SP approach.

This validation is harder than it might first appear. The major problems involve the SP data from the VCE being "clean" in the sense that all the choice sets are observed (even if nothing is purchased), the attributes are standardized, price is clearly exogenous because it is randomly assigned, and the underlying covariate matrix well-behaved, while the RP data has major problems along every dimension. It is worth spelling out those problems. The first is that zero purchase occasions are not clearly defined. That is, in order to make an apples-to-apples comparison, we can only compare choice occasions where positive quantities were chosen in either the RP or SP data. The second involves standardizing attribute levels, which is particularly important for size and what the tuna is packed in, where we only use small and large size cans and packed in oil or water.[8] The third involves an indicator for deals/promotion which is often missing, at least for canned tuna. Visits to shelves of canned tuna at multiple locations and times also suggested that such an indicator even when present is close to useless as a large fraction of canned tuna SKUs had little shelf signs indicating a sale or promotion. So, we don't use these indicators in the IRI RP data and, in parallel, we do not use RP choice alternatives that have money-off coupons.

The second turns out to be much more severe and makes any estimation of count data models using similar RP data highly problematic. We simply do not know what else was on the shelf when the consumer made their canned tuna choice(s).[9] This makes it impossible to use the modelling strategy described below for the RP data which exploits the fact that respondents simultaneously see ten products for which they make volumetric choices.

As a consequence, we fit the simplest model that is compatible with both the RP and SP data, after dropping non-conformable observations in the manner described earlier. This model is a quasi-maximum likelihood truncated Poisson regression model which is a linear function of the log price per ounce. This estimation exercise provides an estimate of the own price elasticity of -0.340, with a 95% confidence interval of [-0.430, -0.249] for the RP choices (N=4,399) and an own price elasticity of -0.373, with a 95% confidence interval of [-0.438, -0.308] for the SP choices (N=13,128). Obviously, these two own price elasticity estimates lie almost on top of each other and in that sense suggest that the SP choices are not inconsistent with those actually being made by IRI panelists with respect to the primary statistic of interest.

Examination of the RP data suggests that the three main brands of tuna are usually available. Adding brand indicators suggests indifference between Starkist (the omitted reference brand) and Bumble Bee for both the RP and SP data. For the RP data, the relative preference for Chicken of the Sea (vs. Starkist) is 8.1% while it is 9.5% using the SP observations, with the difference between the two estimates not statistically significant at any conventional level. We can examine the other attributes, but the estimated magnitude of preference parameters depends critically on what was available from which to choose. Under the assumption that retailers are more likely to offer SKUs with more preferred attribute levels, there should be

[8] Like the amount of cereal in a box, there is some variation in the number of ounces of tuna in a can over time within brand and across brands. Small cans tend to run between almost five ounces and six ounces. We have standardized this to six ounces. Tuna products with four or fewer ounces tend to be pouches rather than cans and appear to represent a different market. Modelling interactions between pouches and small cans might be an interesting extension but is not done here. Large cans tend to run from ten to twelve ounces; we standardized this to twelve ounces. Much larger cans on the order of fifty ounces are sometimes observed, but they seemed to be aimed at a very different market. The number of condiments that can be added to water like mint or sun roasted tomatoes is large. While these can dramatically expand the apparent number of distinct SKUs, we have classified all these as packed in water. We have classified various "marinated in" SKUs as packed in oil.

[9] In empirical work, it is common to infer what other products in a category are available by observing what other consumers in the panel are purchasing and the prices they are paying. While there are no doubt some products for which this provides a reasonable answer to the question of what else is in the choice set, it breaks down quickly in a world where the number of individual SKUs in a category of independent interest is at all large and where stockouts and price changes frequently occur. The difficulty, which holds true even with canned tuna, is that the number of panel participants shopping at a specific store and their frequency of purchase is too small to be able to reliably construct choice sets where errors involving availability and pricing are not a dominant factor in the analysis.

agreement on the signs between the RP and SP models, which is what we see. The small size cans are preferred to large ones, packed in water is preferred to packed in oil, Albacore is preferred to regular tuna, and solid is preferred to chunk. The one place where there is a difference is that the SP data suggest a large coupon effect, while the actual sales data does not, although this may be due to the unreliability of that variable in the RP dataset.

MODEL SPECIFICATION

Our VCE design allows for the estimation of a very rich deeply parameterized model, with a focus on being able to provide own price elasticities for each brand-size combination and a complete set of cross-price elasticities. The availability of individual panelist demographics allowed us to explore their role in driving volumetric choices and understanding price sensitivity. The negative binomial count data model we fit allows deviation from the usual Poisson equality restriction on the mean and variance.

Our model incorporates several specific features which help to appropriately exploit the structure of our VCE. First each of our observations faces multiple choices within a choice set and faces multiple choice sets. We observe multiple choice occasions by our respondents and handle this by allowing for possible correlation between individual-level unobservable components along with robust standard errors, which effectively reduces sample size and prevents artificial inflation of statistical significance levels. Second, the sizeable dimensions of our respondent sample size, choice alternatives, and choice sets, allows us to obtain precise estimates of the nature of preference heterogeneity. We model preference heterogeneity in a way that is likely to make the results more useful for decision making than is the standard practice of allowing for a relatively small number of random parameters. This is done is by using a multilevel mixed-effects specification where many covariates have fixed parameters, which accounts for much of the preference heterogeneity, while allowing other variables to be represented by random parameters and therefore absorb much of the remaining preference heterogeneity.

A novel aspect of our model specification involves a full set of well-behaved own and cross price elasticity terms. This is done by including the log price of each choice alternative in the model as well as the log prices of all the other goods that a respondent also could have purchased in the choice set. Note that this is both similar to what is often done using aggregate data in regression models, but greatly expands the number of cross price elasticities that can be estimated without imposing strong restrictions (in prior empirical work with an appreciable number of competing options) on the relationship between the elasticities (Liu, Otter and Allenby, 2009).[10] Effectively, our specification looks like a mother logit model (McFadden, 1975; Timmermans, Borgers and van der Waerden, 1991) in that we are including the

[10] Chetty (2009) discusses how policy/welfare analysis for marginal changes can generally reliably be undertaken using a small set of reliably estimated reduced form elasticities, which avoids the need for more complex structural estimation which by design depends heavily on unverifiable assumptions. Our VCE guarantees the validity of the reduced form model we estimated for consumers. It does not incorporate any information on how supermarkets or canned tuna producers (who have been accused by the U.S. Department of Justice of colluding in the past) adjust to consumer behavior.

characteristics of the other products as attributes of the product over which the binary buy/no buy decision is being made subject to the constraint that only one good is purchased. While the mother logit model often fits quite well and is an important component of building tests of the adequacy of the conditional logit model specification, it has not typically been used as an operational specification for applied empirical work because the conditional logit model is not a valid RUM when the other products attributes enter directly. That technical problem does not exist with count data models because of separation of preference and scale parameters. The practical problem in RP data is that the prices of different offerings usually are highly correlated and strategically set. Our VCE framework avoids both of those issues.

The next part of the model's structure is standard and involves the product attributes. These appear as indicator variables and include the brand specific constants (except coupon condition). In our model, coupons are represented by two variables, an indicator variable for whether a specific choice option has a coupon and the log of its amount if present (and zero otherwise). This allows us to test whether the coupon presence has an influence distinct from the amount of the coupon and to also look at whether respondents treat the amount of the coupon like a price reduction.[11] The parameters in this part of the model are specified as having random parameters with normal distributions.

Demographic variables include Female, White, Hispanic, and Presence of Children as binary indicators, household size, income (as a set of five categories), and Census Region (as a set of four categories). Each is entered by itself and interacted with its own price elasticity. This allows a wide range in individual-level variation of own price elasticities tied to variation in observable covariates, which underpins the set of brand specific own and cross price elasticities noted above. The essential restriction is that we allow a flexible structure for own price elasticity but impose the restriction that the relative own price elasticities across two respondents are the same. This restriction is not important from the prospective of an individual looking at a shelf of canned tuna and trying to decide what to buy, because the retailer cannot offer different prices to individuals with different characteristics. However, it may be useful to relax this restriction in deciding who to target in promotions or if it is possible to set different (relative) prices across stores whose customer composition differs. In contrast to the product attribute parameters, all coefficients here are fixed. However, we do allow for a random intercept which provides an estimate of the distribution of demand heterogeneity not captured elsewhere.

Finally, we include a set of forty-five random covariance terms and a set of indicators for the specific block of choice sets a respondent saw. These covariance terms pick up preference heterogeneity between pairs of brands, between brands and product attributes, and between the constant term and product attributes. These terms have some similarities to allowing some of off-diagonal terms of the variance-covariance matrix in a random coefficients model to take on non-zero values. While

[11] We did not include the coupon indicator and log coupon value variable in any of our model's random components in order to provide a clean test of how our coupons worked on average. In an exercise targeting coupons, knowledge of how to exploit the variability in coupon response would be useful.

these terms can be difficult to individually interpret, they are a reasonably flexible approach to capturing a first order approximation to a much more complex pattern of underlying substitution relationships. The second is an important control because some blocks of choice options have idiosyncratic components that should not be confounded with the parameters of interest.

Below we present a summary of results from this model. The results should be considered preliminary since additional refinements to the model are needed such as formally incorporating censoring (see Figure 1) of counts above six.

Summary of Preliminary Results

Our modelling objectives were to estimate a complete set of own price elasticities. These estimates, which should be considered preliminary, are shown immediately below in Table 2 where Own_Star12 is the own price elasticity for Starkist 12 oz. cans; likewise Own_Star6 is the own price elasticity for Starkist 6 oz. cans. The other own price elasticities are defined in an analogous way with BB representing Bumble Bee, CofS representing Chicken of the Sea, StoreB representing "store brand," and OtherB represent some other brand. It is important to note that these parameter estimates reflect the relative differences between own price elasticities of brands rather than the actual brand price elasticity, because we allow individual demographics to interact with an individual's generic price elasticity. This component of the estimated model parameters is discussed in more detail below.

Table 2: Own Price Elasticity Estimates

```
                         Robust
      choice | Coef.     Std. Err.   z   P>|z| [95% Conf. Interval]
---------------------------------------------------------------------
Own_Star12   | -3.055376 .1812397 -16.86 0.000 -3.410599 -2.700153
Own_Star6    | -2.977139 .1604504 -18.55 0.000 -3.291616 -2.662662
Own_BB12     | -2.827225 .1489145 -18.99 0.000 -3.119092 -2.535357
Own_BB6      | -3.064267 .2036187 -15.05 0.000 -3.463353 -2.665182
Own_CofS12   | -2.115957 .1306179 -16.20 0.000 -2.371964 -1.859951
Own_CofS6    | -2.721280 .1373983 -19.81 0.000 -2.990575 -2.451984
Own_StoreB12 | -2.169926 .1449513 -14.97 0.000 -2.454025 -1.885827
Own_StoreB6  | -2.616834 .1396330 -18.74 0.000 -2.890509 -2.343158
Own_OtherB12 | -2.827799 .1398006 -20.23 0.000 -3.101803 -2.553794
Own_OtherB6  | -2.916025 .1516987 -19.22 0.000 -3.213349 -2.618701
```

We also estimated 90 cross-price elasticities by including all the prices faced in the other nine volumetric choices available in each choice set. Of the effects estimated, 74 were insignificant at the 0.05 level; they were generally positive and small. Another 16 were significant at the 0.05 level; some were sizeable and could play a potentially important role in pricing decisions. For example, the cross-price elasticity of large size Starkist with large size Chicken of the Sea was estimated to be 0.258 (z=4.15). There was also an interesting price effect where the least expensive per ounce can in the choice set was associated with higher quantities chosen (conditional on all other covariates), which suggests that pricing in both absolute and relative terms can matter at the low end of the cost distribution.

Our model specification also includes estimates for brand-specific constants, where the omitted brand is Starkist. Estimates for the three major tuna brands were surprisingly small. This result is due to most of the action taking place with price elasticities. Here it is worth noting that the ability to disentangle brand-specific parameters and price elasticities is a major strength of the VCE approach.

The estimated parameters of the next part of the model summarize preferences for other product attributes. The results indicate smaller sizes are strongly preferred to larger sizes, tuna packed in water is very greatly preferred to tuna packed in oil, and Albacore tuna was weakly preferred to ordinary (light) tuna. We also find solid is preferred to chunk but the difference is statistically insignificant.

Preferences for whether tuna is packed in oil or water were sufficiently strong that it is uncommon for a single respondent to purchase both in any of the 16 choice sets.[12] This suggests that the design of VCEs can be used to simulate some aspects of a stockout. We found an asymmetric effect where sales of a brand-size combination increase if packed in water if the other size of that brand is packed in oil.

We also estimate the effects of demographic (including spatial) demand drivers. Results for this part of the model revealed that household size was a strong positive predictor of quantity chosen. The presence of children in the household did not matter (conditional on household size), buyers in the South and the West demand larger quantities than those in the Midwest and Northeast, and lower income households buy smaller quantities. Men and women do not differ in their preferences (conditional on controlling for other covariates) including exhibiting the same own price elasticity. White and non-white households do not differ in their preferences, but Hispanics demand more cans. Own price elasticities vary with income category and across the four Census regions.

We estimated random components associated with the brands and sizes and found all 10 variance terms were significant at 0.01 level. Heterogeneity was particularly pronounced for the store brand and any other brand. This is not surprising since these two brand constants lack specificity and may have been interpreted very differently across respondents (although we provided named examples). Preference heterogeneity was relatively small for packed in water versus packed in oil, but quite diffuse for solid versus chunk. The large estimate for the standard deviation of the constant term suggest that there is deep heterogeneity in the demand for canned tuna even after controlling for a sizeable set of observable covariates thought to potentially influence demand. Thirty of the forty-five possible covariance terms between the brands themselves, between the brands and product attributes, and with the constant term are significant at the 0.01 level. The largest covariances in terms of magnitude were between different brand indicators and between the constant term and some attribute indicators. The covariance term between the constant term and the log of own price is particularly large. Our random effects specification allows a rich characterization of heterogeneity, but it causes the model to be computationally difficult to estimate. Finally, looking at the estimate for the negative binomial scale

[12] True lexicographic preferences for all respondents are unlikely as long as there is some within household preference heterogeneity and a willingness by some to tradeoff price against what medium their tuna is packed in.

parameter, equality of the conditional mean and variance is rejected at the 0.01 level in favour of over-dispersion.

The R-square from a regression of the volumetric choices made on the predicted counts for each choice opportunity is 0.72. This is a large improvement over that achieved from any of the suite of the modelling approaches currently in use with data from an earlier (and simpler) version of our VCE with canned tuna (Eagle, Louviere and Islam, 2018). Again, however, the preliminary nature of our results should be emphasized as we are currently exploring the merits of different ways of evaluating the explanatory and forecasting ability of the approach put forward here.

Concluding Remarks

Volumetric choice experiments (VCE) are a natural extension to DCEs that mirror many real-world decisions involving how many units of a good to buy or how many times to undertake an activity in a specified time period. VCEs can be fit using count data models which have well-developed theoretical and statistical foundations. They overcome several long-standing problems with RUM by being able to separately estimate the scale parameter(s) and the parameters of the conditional mean function. This often results in being able to incorporate richer more stable specification of preference heterogeneity into the model. Further, count data models are easily adapted to handle a variety of issues such as censoring and truncation that are often a result of the data collection context.

VCEs can overcome many problems with non-experimental purchase data, avoiding endogeneity problems through randomization of attribute levels and clearly defined choice sets. By simultaneously offering a substantial number of distinct volumetric choices among competing products, we can estimate a complete set of own and cross-price elasticities without the need to restrict the relationship between those elasticities. These elasticities are the key to making efficient decisions about pricing.

Experimental design for volumetric choice is in its infancy, but there are many useful lessons that can be drawn from experience with DCEs. For instance, in the work reported here it would be possible to use the approach developed by Day et al. (2012) to look at whether different segments of our respondents change their price expectations over the course of the choice sets they encounter. In our current work, we have taken the distribution of current canned tuna holdings at home as an unobserved source of heterogeneity. By assigning subsets of respondents to treatments that vary canned tuna holdings or that expose them to different statements about prices likely to be faced in the near future in a manner similar to that done in DCEs, VCEs can be used to shed light on behavioural responses related to these issues.

Comparison between the RP data on actual purchases by IRI Panelists and the behaviour of those Panelists when faced with the canned tuna choices offered in our VCE is less straightforward than it might at first appear. Some issues involve the typical messiness of RP data with a proliferation of SKUs with minor differences and unreliable data fields (if they don’t refer to the physical properties of an SKU).

However, the two largest problems are fundamental: 1) only purchases of positive quantities are observed which can be addressed with considerable loss of information by using only positive purchase quantities from both samples and a truncated count model; 2) (largely uncorrectable) one does not know what else was on the shelf when IRI Panelists made actual purchases, whereas the VCE tells us exactly what else is on the shelf, which is experimentally controlled. Nevertheless, with the weakest assumption that allows identification of the own price elasticity we find what are effectively the same estimate. Slightly stronger assumptions yield very similar estimates for preferences involving the three brands that appear to be typically available. While the magnitude of preferences for attributes like packed in water versus oil cannot be statistically identified without knowing what was available, making the assumption that retailers are more likely to offer products with more preferred attributes suggests that the signs on different non-price attributes should be the same in the RP and SP data. This is empirically the case.

Our preliminary results for the negative binomial regression model estimates using the data obtained using the VCE appear quite promising. The set of own price elasticities produced are reasonable in magnitude. One of the main takeaway messages of our results is that taking both differences in own price elasticities and brand specific constants provides a much richer picture about what is happening in the market than models that only allow for one price effect. Here for instance, we see that for 12 oz. cans the own price elasticity for Starkist is almost 45% higher than for the market leader, Chicken of the Sea, but is only 10% higher for 6 oz. cans. Further, while store brands have price elasticities similar to those of Chicken of the Sea, the 6oz. and 12 oz. store brand constants are much smaller than those for Chicken of the Sea. A complete set of cross price elasticities is produced and these with a few minor exceptions are consistent with theoretical predictions. Moreover, they are generally small in magnitude, but several are large enough to be important considerations in making pricing decisions. Other components of the model paint a rich picture of the underlying preference heterogeneity for different attributes of canned tuna and of the heterogeneity in demographic drivers of demand.

Richard T. Carson

Towhidul Islam

Jordan J. Louviere

References

Allenby, Greg M. (1990), Hypothesis testing with scanner data: the advantage of Bayesian methods. *Journal of marketing research*, 27(4), 379–389.

Day, B., I.J. Bateman, R.T. Carson, D. Dupont, J.J. Louviere, S. Morimoto, R. Scarpa, and P. Wang (2012). Ordering effects and choice set awareness in repeat-response stated preference studies. *Journal of environmental economics and management*, 63(1), 73–91.

Berry, S., J. Levinsohn and A. Pakes. 1995. Automobile prices in market equilibrium. *Econometrica*, 63(4), 841–890.

Cameron, A.C., and P.K. Trivedi (2013). *Regression analysis of count data*, 2nd ed. New York: Cambridge University Press.

Carson, R.T. and W.M. Hanemann (2005). Contingent valuation. In K.G. Mäler and J.R. Vincent, eds. *Handbook of environmental economics vol. 2*, pp. 821–936. Amsterdam: North-Holland/Elsevier.

Carson, R.T., W.M. Hanemann and D. Steinberg (1990). A discrete choice contingent valuation estimate of the value of Kenai King Salmon. *Journal of behavioral economics*, *19*(1), 53–68.

Chetty, R. (2009). Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. Annual Review of Economics, 1(1), 451–488.

Eagle, T.C., J.J. Louviere, and T. Islam (2018). A comparison of volumetric choice models. Presentation at the 20th Sawtooth Software Conference, Orlando Florida, March 7–9.

Goodhardt, G.J. and A.S.C. Ehrenberg (1967). Conditional trend analysis: a breakdown by initial purchasing level. *Journal of Marketing Research*, 4(2), 155–161.

Greene, W. H. (2003). *Econometric analysis*, 5th ed. Upper Saddle River, NJ: Pearson.

Hardt, N. and P. Kurz (2019). How to predict marketplace demand quantities using volumetric choice experiments. Unpublished paper, Fisher College of Business, Ohio State University, July 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418383

Hellerstein, D.M. and R. Mendelsohn (1993). A theoretical foundation for count data models. *American journal of agricultural economics*, *75*(3), 604–611.

Lee, S. and G.M. Allenby (2014). Modeling indivisible demand. *Marketing science*, 33(3), 364–381.

Liu, Q., T. Otter and G.M. Allenby (2009), Measurement of own- and cross-price effects. In V.R. Rao, ed. *Handbook of pricing research in marketing*, pp. 61–75. Northampton, MA: Edward Elgar.

Louviere, J.J., D.A. Hensher and J.D. Swait (2000). *Stated choice methods: analysis and applications*. New York: Cambridge University Press.

Louviere, J.J., T. Ribeiro and R.T. Carson (2016), Volumetric Choice Experiments. Presentation at the 19th Sawtooth Software Conference, Park City, Utah, September 26–30. https://www.sawtoothsoftware.com/download/conf2016/Volumetric%20Choice%20Experiments_9_19_16.pptx

Louviere, J.J. and G. Woodworth (1983). Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data. *Journal of marketing research*, *20*(4), 350–367.

McFadden, D. (1975). On independence, structure and simultaneity in transportation demand analysis. Working paper d7511, Urban Travel Demand Forecasting Project, Institute of Transportation and Traffic Engineering, University of California, Berkeley.

Phlips, L. (1974). *Applied consumption analysis*. Amsterdam: North-Holland/Elsevier.

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. Hoboken, NJ: Wiley.

Schmittlein, D. C., Bemmaor, A. C., & Morrison, D. G. (1985). Why does the NBD model work? Robustness in representing product purchases, brand purchases and imperfectly recorded purchases. *Marketing Science*, *4*(3), 255–266.

Timmermans, H.J.P., A.W.J. Borgers, PJ.H.J. van der Waerden (1991). Mother logit analysis of substitution effects in consumer shopping destination choice. *Journal of business research*, 23(4), 311–323.

Train, K.E. (2009). *Discrete choice methods with simulation*. New York: Cambridge University Press.

Conjoint Meets AI

Peter Kurz
Stefan Binner
BMS Marketing Research + Strategy

Motivation for This Paper

Artificial Intelligence (AI) and Artificial Neural Networks (ANNs) are the “talk of the town”! Most AI applications are located in the area of pattern recognition and big data. However, ANNs have also been used in the area of choice behavior in order to identify preference models (e.g., Bishop, 1995). Examples from the field of market research include models for price elasticities in FMCG and car ownership (e.g., Hensher & Ton, 2000; Mohammadian & Miller, 2002) and various other fields. The major advantage of ANNs is that they can efficiently recognize patterns in the data without being explicitly programmed as to where to look. This key feature of ANNs is called the Universal Approximation Theorem (Hornik et al., 1989) and describes their capability to approximate any Data Generating Process (DGP). However, despite the strong pragmatic appeal of ANNs, they have been criticized for being too much data-driven and theory-poor, in effect presenting the analyst with a black box model of the DGP.

As far as we know, past papers are mostly about how ANNs could be used to derive utility values from conjoint exercises (for example, Belyakov, 2019; Alwosheel, van Cranenburgh & Chorus, 2017). In these papers the conclusion is that hierarchical Bayes models perform as well as or better than ANNs for utility estimation. This is why we decided not to look into ANNs for estimation any more deeply, at least for the moment. New developments in the area of ANNs may change this situation in the future and it may become worthwhile to look into the utility estimation topic once more, but in the meantime HB does an excellent job!

The situation is different in the area of *experimental design* for Choice-Based Conjoint experiments, especially when alternative-specific designs are needed. We haven’t found any literature so far that deals with ANNs and experimental designs. This is the reason why we want to further explore this topic. This is especially true because we encounter weaknesses in everyday work, mostly in the area of generating acceptable experimental designs for complex choice experiments. Calculating a statistically perfect design is an NP-hard problem.[1]

In day-to-day research work, client studies get more and more demanding, the numbers of attributes and levels are constantly increasing, and sample sizes get smaller and smaller. Therefore, in many cases it is not easy to find good experimental designs with the commonly used algorithms. Necessary restrictions and prohibitions on attribute levels (levels that can’t be shown together) often push the limits in finding an appropriate design. Furthermore, most of the experimental designs used in day-to-day research were developed

[1] NP-hardness (non-deterministic polynomial-time hardness), in computational complexity theory, is the defining property of a class of problems that are informally “at least as hard as the hardest problems in NP.” This usually means that we can’t solve the problem with brute-force algorithms because these would need almost endless time to run, except for trivially small cases.

to optimize *aggregate* models of choice behavior (MNL) and are not optimized to estimate heterogeneity in the context of hierarchical Bayes estimation.

DESIGN PRINCIPLES

Designs that are theoretically efficient usually have undesirable empirical properties: level balance, orthogonality, minimal overlap, and some degree of utility balance. A "perfect" design is completely uncorrelated, all levels appear equal times (preferably all two- and three-way combinations appear equally as well) and there is minimal (often meaning no) overlap between the attribute levels shown in one choice task. Such designs are superior from a statistical point of view, but are sometimes very strange to answer for respondents, because of implausible combinations. Choice tasks with a larger number of concepts, where most of the attributes have fewer levels than the number of concepts shown, often result in two nearly identical concepts due to the goal of reaching level balance. Think, for instance, of choice tasks where only the price attribute is varying and all other attribute levels are the same between the concepts. Other choice tasks might show one clearly superior ("dominant") product, so that the respondent's answer is obvious in advance. The variance of the model parameters estimated from a design depends on the actual parameter values, but most design algorithms assume equal preference weights in calculating efficiency (even with unknown preferences, attributes often have a natural ordering such as "mild," "moderate," and "severe," or a lower price which should be preferred over higher ones). This is the reason why classical designs often result in dominated pairs, where all the attribute levels of one alternative are better than the attribute levels of another alternative. Such choices provide no real preference information, even though they may be included in a theoretically efficient design.

Some researchers have concluded that a certain amount of utility balance (having alternatives about equally attractive overall) is needed, both to avoid odd choice tasks and to enhance efficiency (Huber & Zwerina, 1996). Rich Johnson, Joel Huber, and Bryan Orme (2005) conducted some practical experiments and showed that while utility balance is theoretically good, it usually doesn't result in improvements in empirical studies. This finding is attributed to the idea that utility balanced designs result in choice tasks with concepts that are more nearly equal in attractiveness and therefore make the choice tasks harder for respondents, causing more fatigue and/or random error. Also, creating utility balance requires prior information; if we are able to create perfect utility balance, we already know the answers!

We agree with both papers. On the one hand, no utility balance is bad for respondents, because they may see trivial choice tasks where the answer is clear to everyone. On the other hand, high utility balance may lead to too-difficult choice tasks. The net conclusion is, no utility balance at all and too high utility balance are both bad for empirical studies, but some degree of utility balance does help achieve better designs.

Another frequent concern when building statistically perfect designs is including implausible attribute combinations. An orthogonal design might include all possible combinations, including ones that are not possible in reality. Imagine an iPhone with an Android operating system. Eliminating these implausible combinations results in a design that is no longer orthogonal or level balanced.

Further problems constructing experimental designs often appear when line-pricing or a pricing system from the client has to be taken into account. Pricing systems could be such that if Coke increases in price, line-pricing for other soft drinks of the CocaCola company (e.g., Sprite) need to increase similarly Or, it might be a rule that a 10% price increase in sparkling water should mean a 30% increase for energy drinks. Such pre-conditions and relationships also lead to problems with level balance and orthogonality of the design.

How Do AI and ANNs Work?

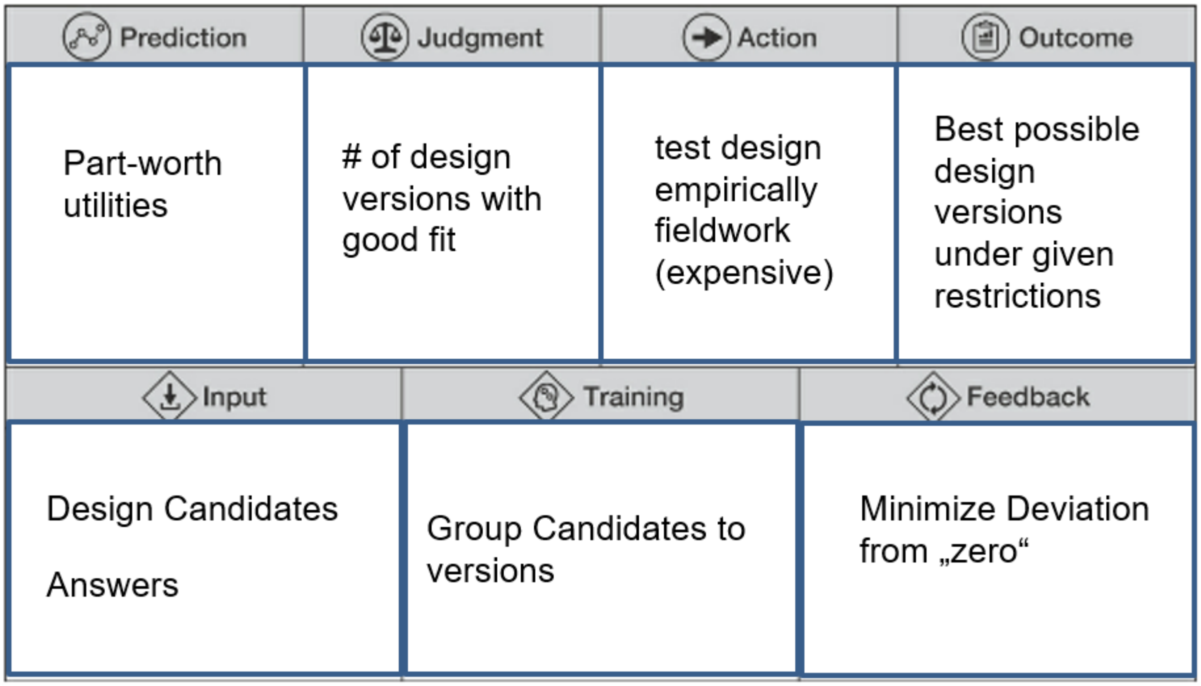
Taking the topics mentioned above into account, an AI-based design generation process has to find a nearly perfect experimental design and to minimize the statistical and measurement error. Such an AI approach should accommodate all practical needs like prohibitions, excluding implausible combinations and unavailable products, taking pricing systems into account, and including at least some utility balance to make the choice tasks more realistic. But, it must do this while giving up as little as possible of the desirable statistical properties (e.g., orthogonality, level-balance, overlap). From a statistical point of view, this means that we are looking for design versions in which purely random answers result in all estimated part-worth values being zero, or nearly so, and therefore an RMSE being as small as possible (standard errors close to zero). This sounds like more of a challenge than it actually is, because exactly such requirements are the strengths of ANNs.

What is the starting point for an ANN-based design approach?

- We know the answers to test (simply random answers).
- We know which prohibitions and pricing rules we must take into account.
- It is relatively easy to generate a large number of synthetic datasets to train our ANN's.
- We have a clear objective and "loss function": minimizing the standard errors of the part-worth coefficients.

Before building an AI tool it's necessary to decompose the workflow into separate tasks. Agrawal, Gans & Goldfarb (2018) suggest the use of an "AI canvas" (Figure 1) that helps to decompose the machine learning problem into tasks:

Figure 1: The AI Canvas



Prediction

Part-worth utilities

Judgment

# of design versions with good fit

Action

test design empirically fieldwork (expensive)

Outcome

Best possible design versions under given restrictions

Input

Design Candidates

Answers

Training

Group Candidates to versions

Feedback

Minimize Deviation from „zero“

Source: Agrawal, A.; Gans, J.; Goldfarb, A. (2018): Prediction Machines.

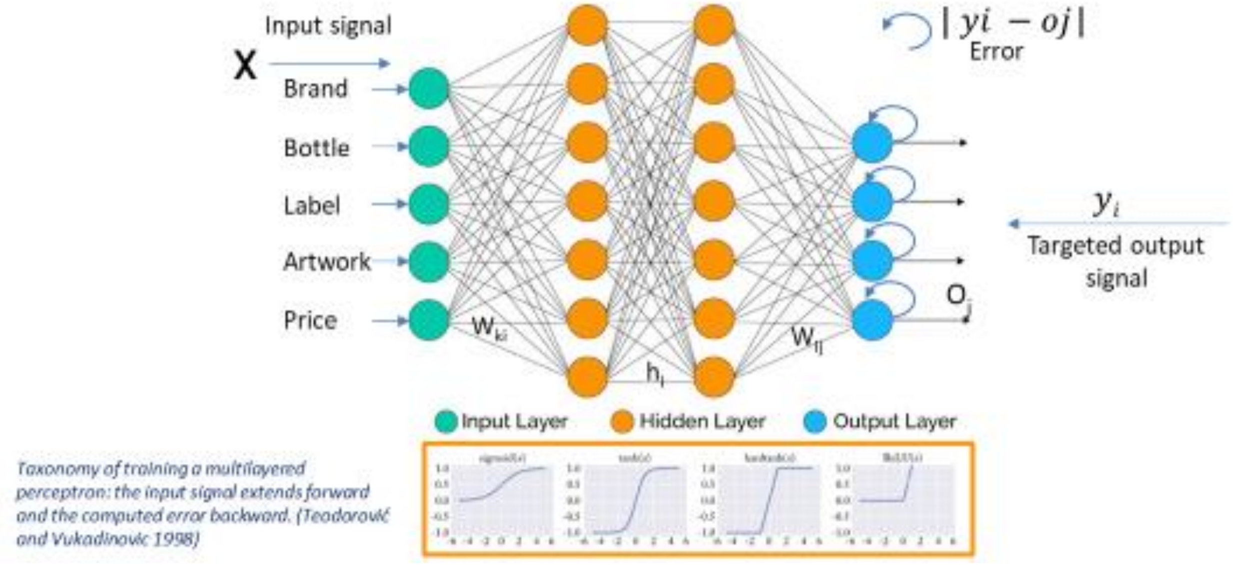
The first task is to **predict** the part-worths, so we need some process within our AI tool to estimate part-worth utilities. This could be done with the Softmax procedure implemented in the Keras[2] AI framework. Softmax is an ANN type that realizes multinominal logit calculations. The second task, **judgment**, determines how many good design versions we find in our input data, based on a maximum loss we will accept from the loss-function (standard error). The (third) **input** task involves generating the possible design versions. A design version is a complete set of choice tasks for one respondent. The candidate design versions can be produced either by fully random design generation or some specialized design algorithm. The (fourth) **training** task involves grouping design versions to make complete experimental designs that minimize the loss. The (fifth task) **feedback** loop is simply minimizing the loss function which means that we want to have part-worths with standard errors as small as possible. Information on the errors is sent back to the training task so it can try again to group the design versions into a different complete experimental design that results in a smaller loss. The (sixth task) **outcome** of this AI tool is the best possible experimental design (or at least, one very close to the best) under the given restrictions.[3] Finally, the (seventh) **action** task is to use the design in empirical studies and compare it to other designs used under similar conditions. Action is the most expensive step in the AI canvas. All the other steps only need computational power and can be done in the lab; action must be in the real world.

[2] Keras is a Python-based package for neural networks, accessible through R as well.

[3] What we call a “choice task” is the single question shown to a respondent. The choice task consists of a number of concepts. What we call a “version” or “design version” is the full set of choice tasks shown to one respondent, typically numbering 8 to 15. The complete experimental design consists of a defined number of versions, usually 20 to 40.

Since ANNs can be used to approximate any data generating process due to the universal approximation theorem, it should not be a problem to find an appropriate design. But, in order to find this solution ANNs need a large amount of data and a properly defined loss function. As shown above, we can easily fulfill both requirements: we can generate a large number of versions and answers and train the ANN to solve the problem by minimizing the loss function (small standard errors) and identify the optimal experimental design (combination of different versions) need for a complex choice experiment based on our input.

Figure 2: How ANNs Work



Taxonomy of training a multilayered perceptron: the input signal extends forward and the computed error backward. (Teodorović and Vukadinovic 1998)

Source: **Teodorović, D., & Vukadinovic, K**. **(1998):** Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach.

ANNs can be described as weighted directed graphs (Figure 2), where the nodes (colored circles) are the “neurons” and the connection lines between the neuron outputs and neuron inputs can be characterized by the targeted edges with weights (W_{ki}). Figure 2 is a simplified example of the real network we used. Our input is much more complex, one neuron for each attribute level * # of concepts * # number of choice tasks * # of versions needed for the experimental design + the synthetic answers for each choice task * number of synthetic respondents. As this is way too complex for an illustration, we show only one neuron for each attribute in Figure 2. The ANN collects the input signal from the external world in the form of a vector with binary information. These inputs are then mathematically defined by the notations x(n) for every n number of inputs, where n is the number of versions.

To make this more concrete, if we have 5 attributes with 3 levels each and 1 with 4 levels, for 19 total levels; 4 concepts per choice task; 12 tasks per version; and 20 versions desired for the complete design, our input layer would have 19 * 4 * 12 * 20 = 18,240 input neurons, each receiving a 0 or 1 input depending on whether a particular level applied to a particular concept in a particular task in a particular version. In addition, with 1,000 synthetic respondents, we would have 4 * 12 * 20 * 1000 input neurons indicating which of the four concepts were chosen in each task in each version by each simulated respondent.

Each of the inputs is then multiplied by its corresponding weights (these weights are determined by the training process used by the artificial neural networks to solve a specific problem). In common terms, these weights represent the intensity of the interconnection among neurons inside the ANN. All the weighted inputs are summed up inside the ANN (this could be seen as another artificial layer of neurons).

The "activation function" of each hidden or output neuron is some transformation of the weighted sum of the input values. It may be sigmoidal, or linear, or a step function, among other possibilities, as suggested by the bottom panel of Figure 2. The "softmax" neuron that implements logit choice is one specialized type of neuron. To understand the architecture of an ANN, we need to understand what components the neural network contains. A typical ANN includes a large number of artificial neurons which are units arranged in a series of layers. Let us take a closer look at the different layers available in an ANN:

Input Layer

The input layers contain those artificial neurons (units) which directly receive input from the outside world (input data), in our case, the design versions and answers.

Output Layer

The output layers consist of units that react to and reflect the information that is fed into the ANN. How closely their outputs (for us, part-worths) correspond to the desired ones (near-zero part-worths) reflects whether the ANN has learned any task well or not.

Hidden Layer

The hidden layers are located between the input layers and the output layers (not interacting with the outside of the ANN). The only job of a hidden layer is to convert the input into some meaningful form that the next layer can use in some way.

Most ANNs are completely interconnected, which means that all neurons in one layer are connected to all neurons in the next layer, leaving nothing unconnected. This allows a complete learning process. Learning occurs when the weights inside the ANN get updated after each new iteration.

In order to generate the input for training the ANNs we developed R code that generates large numbers of versions which are used as input for the first layer. The code generates versions based on a modula operation, by simply cycling thru all attribute-level combinations. In a second step the code deletes any tasks that fail any of our defined conditions: prohibitions, implausible combinations between concepts, violation of line-pricing, non-conformity with the needed pricing systems, violation of the desired amount of utility balance, and so on. To ensure that we produce enough possible versions, we use an oversampling strategy to bring in more versions with attribute-level combinations that are "relatively rare" because of the various restrictive conditions. We run these R scripts until we reach a large number of possible versions (typically a million or so) and then use those possible versions as input for the ANN.

The second part of the training data is the answers to the choice tasks. We simply generate random choices of concepts in each choice task. It is crucial to use a perfect random distribution and avoid all positional effects. However, if the eventual real-world experiment will include a None option, it is important to define a realistic frequency of None answers and include them in the training data. Otherwise it is not possible to estimate the effect of None answers in the design, meaning that the efficiency will always be overestimated because of the assumption that all choice tasks will be answered on a forced-choice basis. Therefore, it is essential to have good estimates of real-world answering behavior, which is not always easy.

In order to train our ANNs we generated 1,000 synthetic respondents each answering all 15-20 choice tasks of a version. We replicated this for all of our 1,000,000 "design versions" that are used to build up the ANN. This procedure gave us sufficient information to estimate part-worth utilities at aggregate level and to have a large enough set of versions to stabilize the training of the ANNs.

Loss Function

A key task is to define the loss function. In the case of Choice-Based Conjoint exercises, it is simply the deviation from zero of the estimated part-worth utilities. The training process iteratively generates weights for the ANN layers which minimize the deviation from zero for the estimated part-worth utilities. As we want no single biased attribute levels, we minimize the deviation for all single attribute levels (parameters to estimate), so that we result in a loss-function combining the overall standard error and the standard error for each single parameter.

The proportion of the two components (overall and individual standard error) can either be a fixed ratio (as we have used in the example 70:30) or varied due to an optimization function (which introduces another layer). If we can get exact "zero"-values for all part-worth estimates, the design fulfills all design considerations. Deviations from zero cause or reflect weaknesses in at least one of the design considerations. Most often, orthogonality is violated (attribute levels are correlated). However, we know from the literature it is acceptable to give up some orthogonality in order to derive useful designs, especially when estimating multinomial logit models. Nevertheless, the better we meet the design considerations, the better our empirical study results.

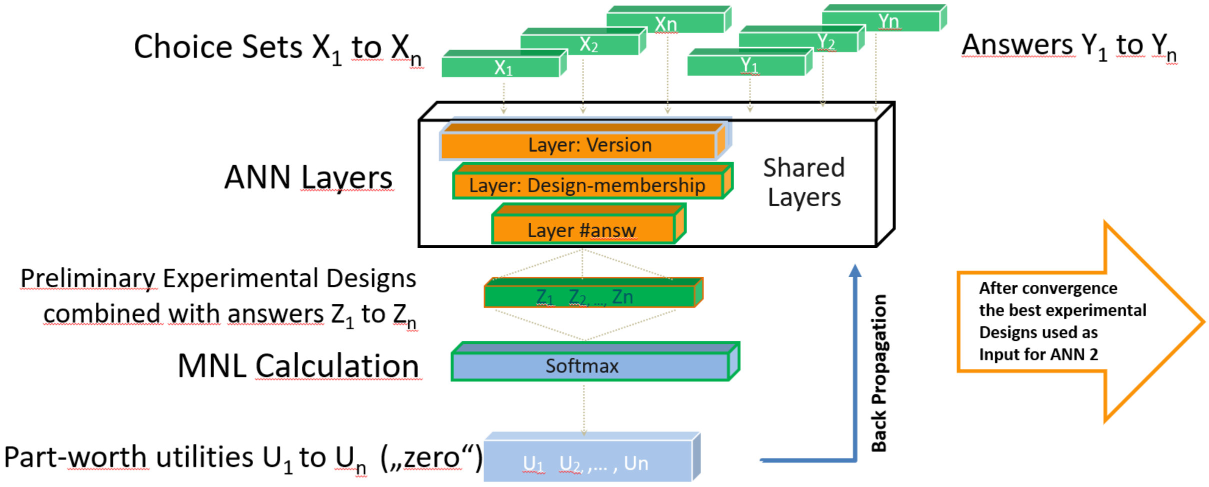
For experimental designs where it is possible to fulfill all design considerations (it is often not possible, due to prohibitions that prevent orthogonality, for example), we can prove that ANNs are able to minimize the loss function to exactly zero values for all estimates.

Training of the ANNs

Finally, we have to train the ANNs. Our experiment with different ANN strategies showed best results, if we train two ANNs: ANN1 for the experimental design that best fits at the aggregate level and ANN2 to optimize the resulting experimental designs for individual utility estimation.

For ANN1 we use the randomly generated versions and random answers to the choice tasks in them as input. Then we train the layers to find the optimal experimental design to estimate the aggregate MNL. This is done using a standard way of training ANNs called "backpropagation," an iterative process that adjusts the weights of the neurons to minimize the loss. After convergence is reached, ANN1 is able to produce the best experimental design possible, for later field work. For the complete experimental design we use a sufficient number of versions, so that different respondents see different sets of choice tasks. From our experience we usually end up with 20 to 40 different versions for a complete experimental design that is optimal for fielding. Up to this point, the design is only tested on an aggregate level, meaning we have used aggregate logit models to define the loss function being minimized.

Figure 3: ANN1 — Design Optimized for *Aggregate* Utility Estimation



Choice Sets X_1 to X_n
Xn
X2
X1
Yn
Y2
Y1
Answers Y_1 to Y_n
ANN Layers
Layer: Version
Layer: Design-membership
Layer #answ
Shared Layers
Preliminary Experimental Designs combined with answers Z_1 to Z_n
Z1 Z2 ,.., Zn
MNL Calculation
Softmax
Part-worth utilities U_1 to U_n („zero")
U1 U2 ,... , Un
Back Propagation
After convergence the best experimental Designs used as Input for ANN 2

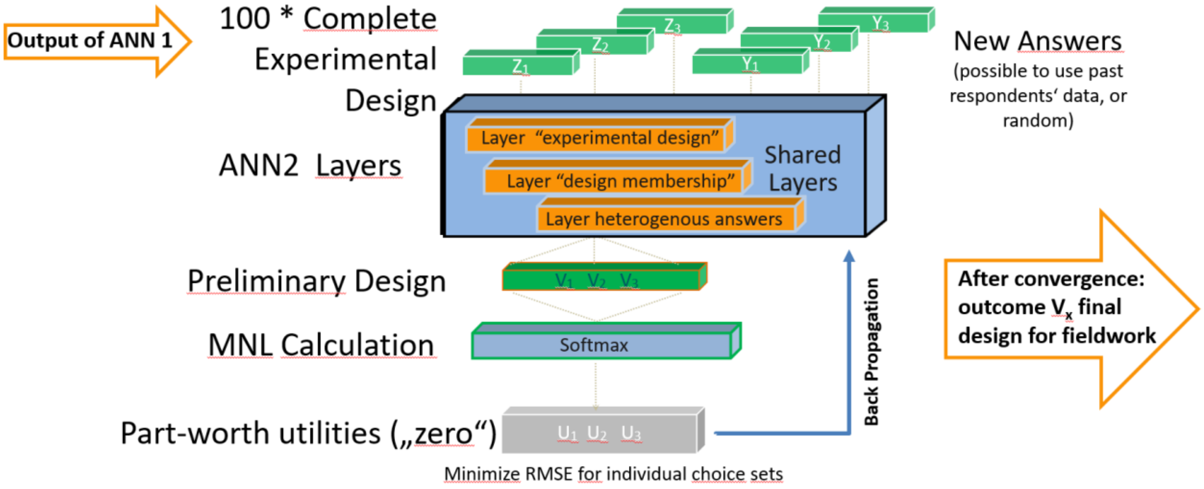
Flow of the first ANN: Input, Layers and Softmax Layer for MNL Calculation

ANN2 is trained to find the best experimental design for individual utility estimation (via HB) based on a sufficient number of possible experimental designs generated by ANN1. The input data are now complete experimental designs (the results from ANN1).[4] These designs are answered with new randomly generated answers. But at this stage we add some heterogeneity and eventually response error to the answers and train the ANN2 to minimize the deviation for each experimental design. To estimate on the individual level, the versions from each of the experimental designs are answered by 1,000 synthetic respondents, which are generated according to the assumed heterogeneity. For our training data we used 100 experimental designs, of 20 versions each, generated by ANN1 and answers to each of them by 1,000 synthetic respondents.

As one can see, in this phase we need assumptions about the real heterogeneity in the population. In future work we would like to use real data from past studies at this stage to generate synthetic respondents' answers, or at least generate answers based on knowledge of heterogeneity from past studies or markets.

[4] To implement the second step we uses a number of possible "nearly optimal"experimental designs from ANN1. Usually we find not just one perfect experimental design, but a group of possible experimental designs. If we end up with only one one perfect experimental design, we have to lower the restrictions of the loss-function a little and re-run to get more design options.

Figure 4: Design Optimized for *Individual* Utility Estimation


Output of ANN 1
100 * Complete Experimental Design
Z1
Z2
Z3
Y1
Y2
Y3
New Answers
(possible to use past respondents' data, or random)
ANN2 Layers
Layer "experimental design"
Layer "design membership"
Layer heterogenous answers
Shared Layers
Preliminary Design
V1 V2 V3
MNL Calculation
Softmax
Part-worth utilities („zero")
U1 U2 U3
Back Propagation
After convergence: outcome Vx final design for fieldwork
Minimize RMSE for individual choice sets

Flow of the second ANN: Input (results from ANN1), Layers and Softmax Layer for individual MNL Calculation

After we have trained the two ANNs, we are able to use the two networks to generate the best possible design versions for fieldwork. The ANNs have recovered the data-generating process which lies behind the design versions and the answers.

Bear in mind that the ANNs automatically structure the layers and calculate the weights, so there is no "programming" needed, the only information we used in the setup is the difference between versions and answers, and in ANN2 an indicator to which experimental design the versions belong. However, the resulting ANNs are "black boxes," so we don't have a chance to analyze them and see why some selections are superior to others and why we end up with those final design versions. This immediately shows the need of empirical testing on how good the results really are in reality!

Simulation Results

We ran hundreds of synthetic datasets to explore how well ANNs are able to generate ideal experimental designs when the underlying DGP (known utilities, Gumbel error as in standard MNL) is known to the analyst. We focus on standard criteria for good experimental designs like orthogonality, level balanced overlap, and utility balance (see Huber & Zwerina, 1996).

Table 1: Experiment Factors and Factor Levels

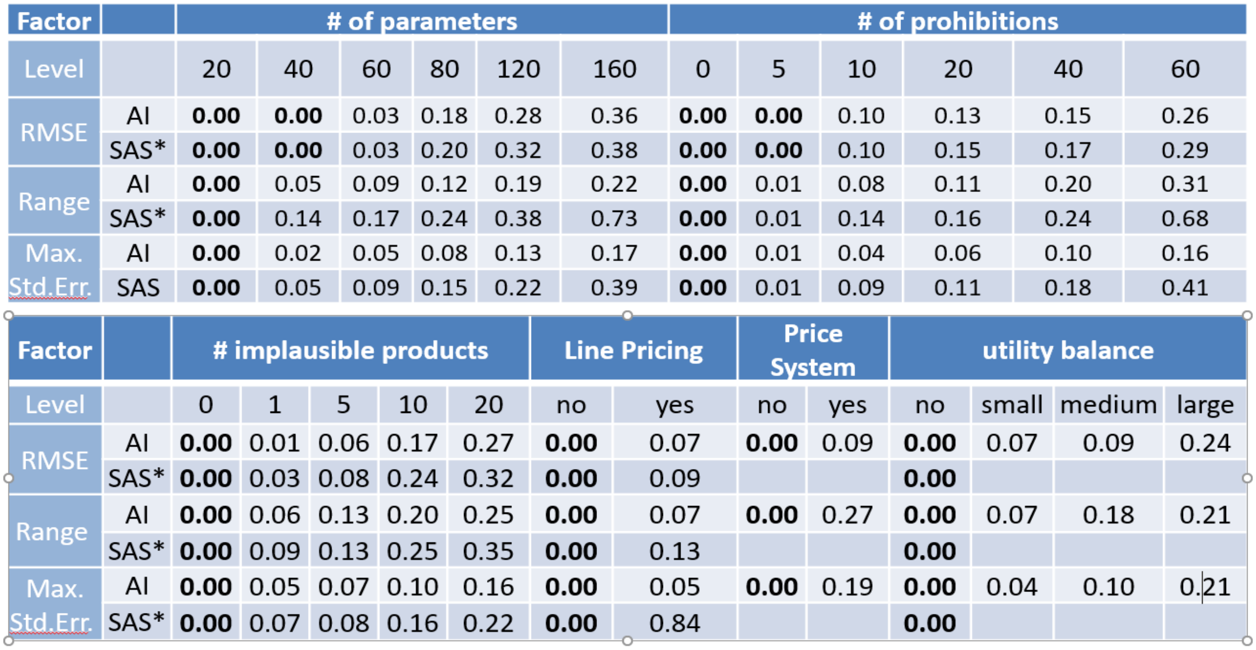
#	Factor	#Factor levels	Factor levels
1	# of parameters	6	20;40;60;80;120;160
2	# of prohibitions	6	0; 5, 10, 20, 40, 60,
3	Implausible products	5	0, 1, 5, 10, 20,
4	Line-Pricing	2	yes/no
5	Price System	2	yes/no
6	Utility balance	4	no, small, medium, high

There are 2.880 experimental conditions for the 6 experimental factors.
The number reduces slightly, cause some of the conditions could not be combined (for example 20 parameters with 120 prohibitions)

We used 6 experimental factors (Table 1) and varied them with 2 to 6 factor levels. For each of the combinations we generated experimental designs based on our two-stage ANN approach and answered them with synthetic respondents. As a comparison, we produced designs with the SAS Macros (Kuhfeld, 1996) with the same experimental factors and answered them with the same synthetic respondents. We chose SAS because it is possible to script the macros in SAS syntax to generate the over 2,000 designs automatically. In this experimental setting, we know the real utilities of each respondent and can show how well we reproduce the preferences of the artificial respondents.

To compare our different designs we used the root mean squared error (RMSE) over all part-worth utilities, the maximum standard error, and the range of the standard errors as criteria. Table 2 displays the results of this comparison for all experimental factors.

Table 2: Results from Simulation Study



Factor		# of parameters						# of prohibitions					
Level		20	40	60	80	120	160	0	5	10	20	40	60
RMSE	AI	**0.00**	**0.00**	0.03	0.18	0.28	0.36	**0.00**	**0.00**	0.10	0.13	0.15	0.26
	SAS*	**0.00**	**0.00**	0.03	0.20	0.32	0.38	**0.00**	**0.00**	0.10	0.15	0.17	0.29
Range	AI	**0.00**	0.05	0.09	0.12	0.19	0.22	**0.00**	0.01	0.08	0.11	0.20	0.31
	SAS*	**0.00**	0.14	0.17	0.24	0.38	0.73	**0.00**	0.01	0.14	0.16	0.24	0.68
Max. Std.Err.	AI	**0.00**	0.02	0.05	0.08	0.13	0.17	**0.00**	0.01	0.04	0.06	0.10	0.16
	SAS	**0.00**	0.05	0.09	0.15	0.22	0.39	**0.00**	0.01	0.09	0.11	0.18	0.41

Factor		# implausible products					Line Pricing		Price System		utility balance			
Level		0	1	5	10	20	no	yes	no	yes	no	small	medium	large
RMSE	AI	**0.00**	0.01	0.06	0.17	0.27	**0.00**	0.07	**0.00**	0.09	**0.00**	0.07	0.09	0.24
	SAS*	**0.00**	0.03	0.08	0.24	0.32	**0.00**	0.09			**0.00**			
Range	AI	**0.00**	0.06	0.13	0.20	0.25	**0.00**	0.07	**0.00**	0.27	**0.00**	0.07	0.18	0.21
	SAS*	**0.00**	0.09	0.13	0.25	0.35	**0.00**	0.13			**0.00**			
Max. Std.Err.	AI	**0.00**	0.05	0.07	0.10	0.16	**0.00**	0.05	**0.00**	0.19	**0.00**	0.04	0.10	0.21
	SAS*	**0.00**	0.07	0.08	0.16	0.22	**0.00**	0.84			**0.00**			

The differences between AI and SAS become larger the more complex the designs are: the more prohibitions and implausibilities that have to be taken into account, the better the ANN-based versions perform compared to the SAS designs. Furthermore, we can see that a high degree of utility balance harms the designs. This finding is in line with the literature: a small amount of utility balance is OK, but higher utility balance makes the designs worse. Note: in the case of synthetic data, the problem with utility balance is not respondent burden or error. With higher utility balance we simply produce larger deviations from orthogonality, and for complex designs with lots of prohibitions we also violate level balance. In the real world, it's likely that respondents answering behavior and higher burden, as Rich Johnson encountered in his studies, eventually add further to these errors.

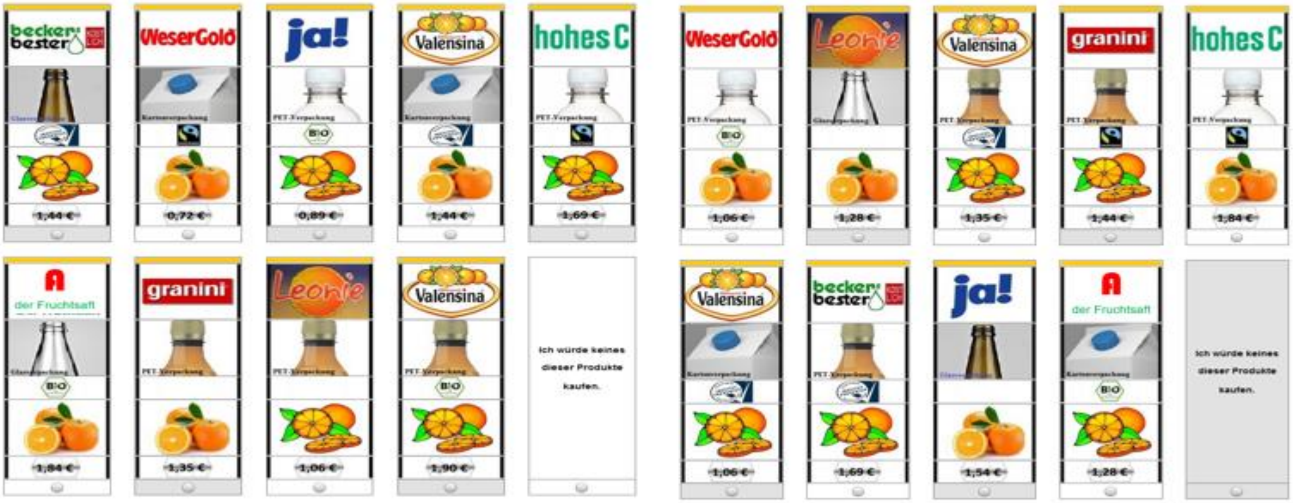
The same pattern can be seen in hit rates and shares. The AI-based designs always have lower RMSE than the classical SAS-based designs. For the synthetic data, we can conclude that the black-box works and the ANNs are well-trained to produce appropriate designs and can handle much more complex design restrictions, compared to designs based on the classical algorithms (the SAS designs). The results for hit rates show that the ANN-based designs are better at capturing heterogeneity than the classical designs. This leads to the conclusion that the training effort for the second ANN pays off.

Empirical Studies

For empirical testing of ANN performance with real data we conducted two studies with split cell designs: one study about orange juice and one about chocolate bars. In each study, one cell used ANN-based designs, and the other used designs from the SAS macros.

The orange juice study was conducted in Germany in December 2018 and January 2019 with respondents who bought orange juice in the last month, aged between 16 and 65 years. The sample was n=1,010 and n=1,005 for the two design splits. The objective of this study was to optimize the bottle type, the quality of juice, and the packaging artwork, and to get insights into the impact of quality labels like "organic" or "fair trade" (Figure 5). The design complexity comes from the problem that not all brands can produce all package types or all different qualities. Depending on the quality and pack type, not all prices could appear with all juices.

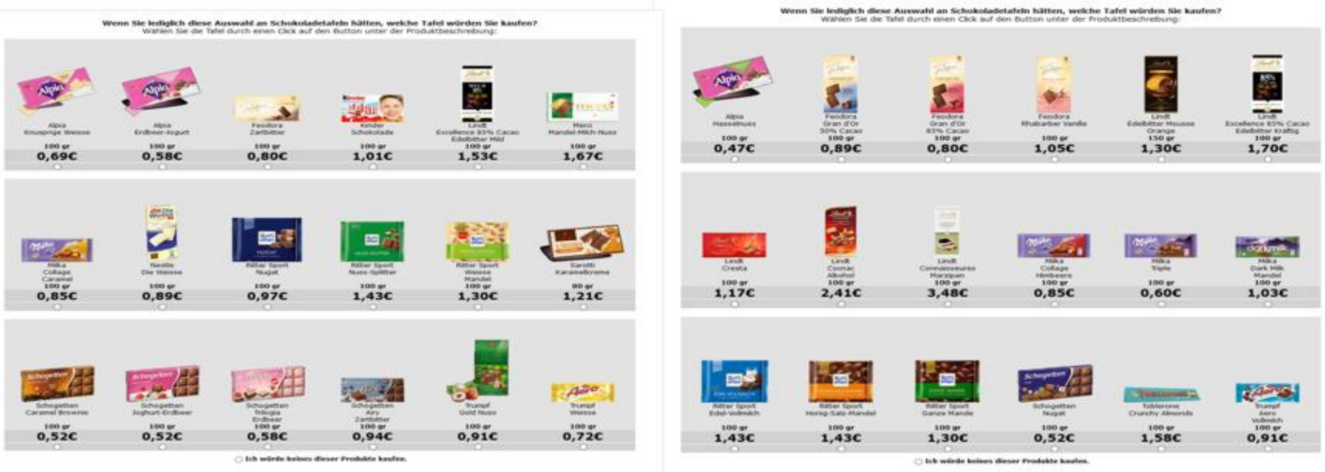
Figure 5: Choice Tasks from the Orange Juice Study



Orange Juice study Germany 2018/19; n=1010/1005

The second study, on chocolate bars, was also conducted in Germany, in August 2018 and July 2019 with two samples of 605 and 608 category buyers (at least once in the last month) and aged between 16 and 65 years. The objective of this exercise was to optimize the assortment and pricing, add new flavors to the market, and have an optimal differentiation from what the competitors offer. The design complexity was caused by the fact that not all brands can offer all flavors, some brands use very special ingredients that are branded and cannot be used by other brands, and some brands produce special chocolate variants with special cacao that cannot be shown with other brands. In addition, the assortment size is very different between the competitors and some ingredients are much more expensive than others (Figure 6).

Figure 6: Choice Tasks from the Chocolate Bar Study



Wenn Sie lediglich diese Auswahl an Schokoladetafeln hätten, welche Tafel würden Sie kaufen?
Wählen Sie die Tafel durch einen Click auf den Button unter der Produktbeschreibung:

Alpia Knusprige Weisse 100 gr 0,69€
Alpia Erdbeer-Jogurt 100 gr 0,58€
Feodora Zartbitter 100 gr 0,80€
Kinder Schokolade 100 gr 1,01€
Lindt Excellence 85% Cacao Edelbitter Mild 100 gr 1,53€
Merci Mandel-Milch-Nuss 100 gr 1,67€
Milka Collage Caramel 100 gr 0,85€
Nestlé Die Weisse 100 gr 0,89€
Ritter Sport Nugat 100 gr 0,97€
Ritter Sport Nuss-Splitter 100 gr 1,43€
Ritter Sport Weisse Mandel 100 gr 1,30€
Sarotti Karamellcreme 80 gr 1,21€
Schogetten Caramel Brownie 100 gr 0,52€
Schogetten Joghurt-Erdbeer 100 gr 0,52€
Schogetten Trilogia Erdbeer 100 gr 0,58€
Schogetten Airy Zartbitter 100 gr 0,94€
Trumpf Gold Nuss 100 gr 0,91€
Trumpf Weisse 100 gr 0,72€

○ Ich würde keines dieser Produkte kaufen.

Wenn Sie lediglich diese Auswahl an Schokoladetafeln hätten, welche Tafel würden Sie kaufen?
Wählen Sie die Tafel durch einen Click auf den Button unter der Produktbeschreibung:

Alpia Haselnuss 100 gr 0,47€
Feodora Gran d'Or 50% Cacao 100 gr 0,89€
Feodora Gran d'Or 85% Cacao 100 gr 0,80€
Feodora Rhabarber Vanille 100 gr 1,05€
Lindt Edelbitter Mousse Orange 150 gr 1,30€
Lindt Excellence 85% Cacao Edelbitter Kräftig 100 gr 1,70€
Lindt Cresta 100 gr 1,17€
Lindt Cocnac Alkohol 100 gr 2,41€
Lindt Connaisseurs Marzipan 100 gr 3,48€
Milka Collage Himbeere 100 gr 0,85€
Milka Triple 100 gr 0,60€
Milka Dark Milk Mandel 100 gr 1,03€
Ritter Sport Edel-Vollmilch 100 gr 1,43€
Ritter Sport Honig-Salz-Mandel 100 gr 1,43€
Ritter Sport Ganze Mandel 100 gr 1,30€
Schogetten Nugat 100 gr 0,52€
Toblerone Crunchy Almonds 100 gr 1,58€
Trumpf Aero Vollmilch 100 gr 0,91€

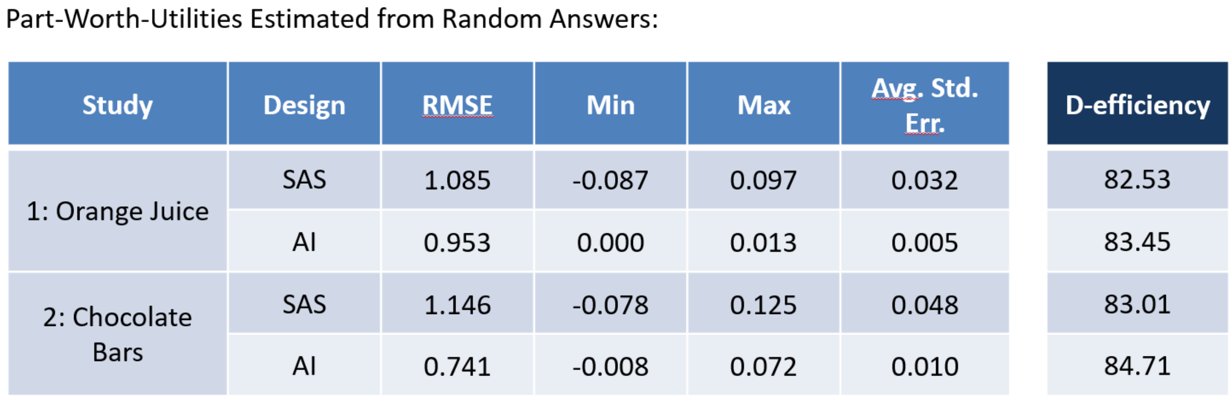
○ Ich würde keines dieser Produkte kaufen.

Chocolate bar study Germany 2018/19; n=605/608

We expected that if we generated the experimental designs for the two studies with the ANN-based approach, it would be possible to improve level balance for one-way and two-way frequencies, as well as orthogonality. Furthermore, we used the ANN2 to add some utility balance for the price attribute, so that lower prices for the same type of chocolate bars are preferred. Due to the limited number of empirical test cells we weren't able to test whether a specific amount of level overlap and/or adding special tasks to estimate interaction effects would pay off. However, ANNs could easily address such criteria.

First, we compared the designs for the two studies, again with purely random answers. With this initial test we could show that the AI-based designs have a smaller RMSE, smaller standard deviation and smaller standard errors. Even though our ANNs were not trained to optimize D-efficiency, they beat the SAS designs in the area of their strength, D-efficiency, as well (Table 3).

Table 3: Comparison of the Two Design Approaches



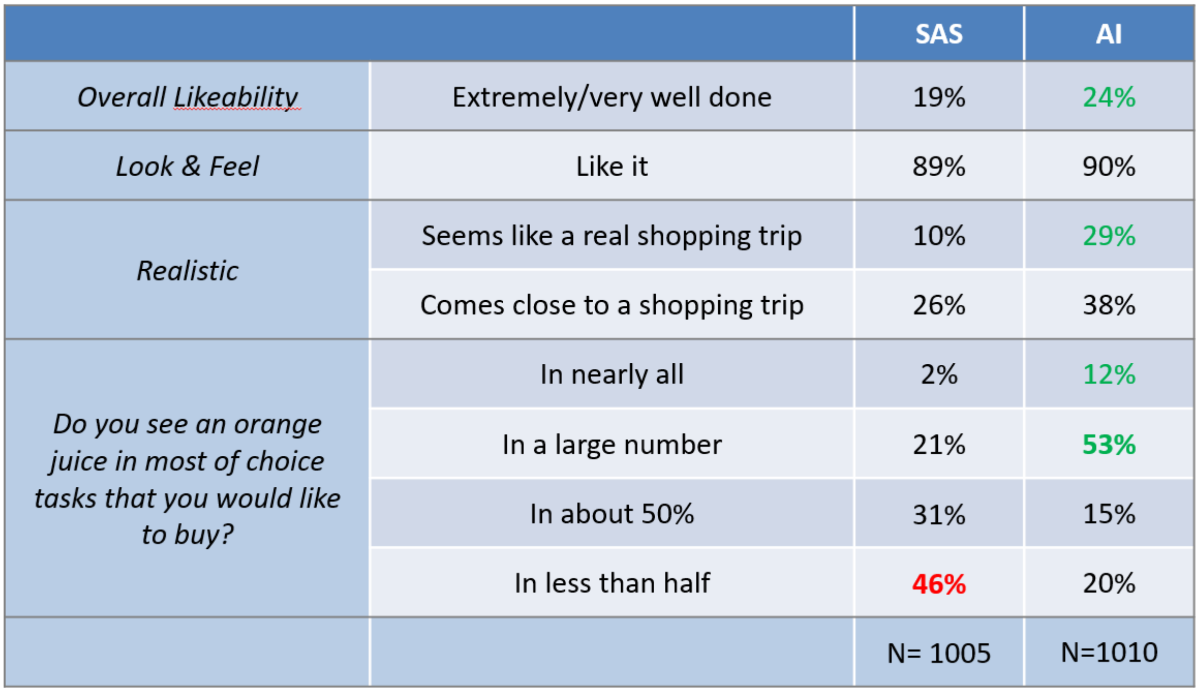
Part-Worth-Utilities Estimated from Random Answers:

Study	Design	RMSE	Min	Max	Avg. Std. Err.	D-efficiency
1: Orange Juice	SAS	1.085	-0.087	0.097	0.032	82.53
	AI	0.953	0.000	0.013	0.005	83.45
2: Chocolate Bars	SAS	1.146	-0.078	0.125	0.048	83.01
	AI	0.741	-0.008	0.072	0.010	84.71

Orange Juice Germany 2018/19; n=1010/1005; Chocolate bar 2018/2019; n=605/608

In the orange juice study, we added additional feedback questions about respondents' experience when doing the exercise. In Table 4, we can see slight improvements in likeability, but even higher gains for the AI-based designs concerning the realism of the exercise compared to a real shopping trip. But more important is the dramatic increase in the number of choice tasks where respondents can make a choice, because they see a product that they would really like to buy. A real advantage of the ANN-based designs is the number of meaningful choice alternatives for respondents. Answers of respondents should be more realistic and meaningful if they really see products they want to buy and don't have to choose products they would never really consider buying in most of the tasks.

Table 4: Results from the Feedback Questions, “Orange Juice” Study

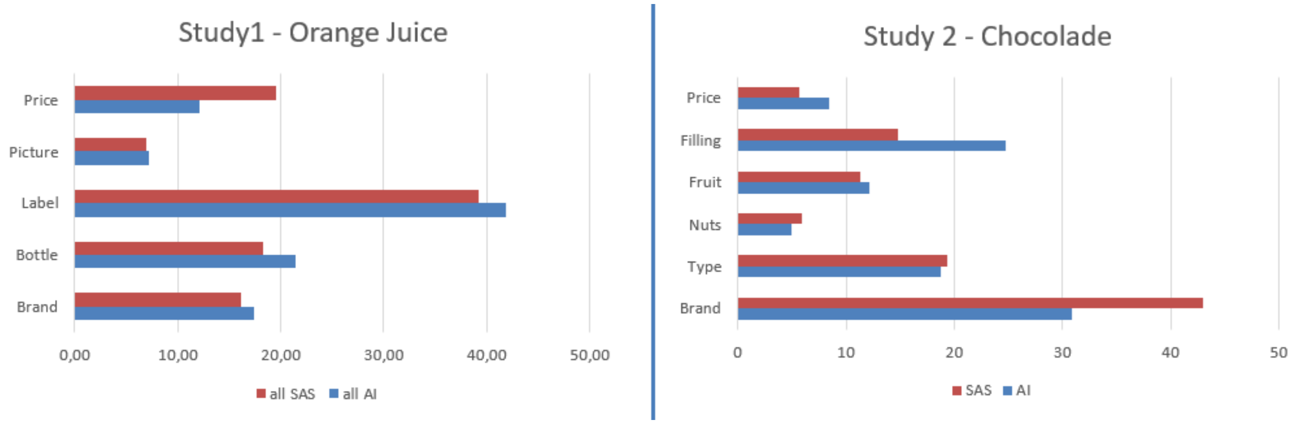


		SAS	AI
Overall Likeability	Extremely/very well done	19%	24%
Look & Feel	Like it	89%	90%
Realistic	Seems like a real shopping trip	10%	29%
	Comes close to a shopping trip	26%	38%
Do you see an orange juice in most of choice tasks that you would like to buy?	In nearly all	2%	12%
	In a large number	21%	**53%**
	In about 50%	31%	15%
	In less than half	**46%**	20%
		N= 1005	N=1010

Orange Juice study Germany 2018/19; n=1010/1005

A second finding is that the importances of the attributes are different between the two design strategies. In Figure 7, we can clearly see that, especially if prohibitions are used, design weaknesses influence the attribute importance. In the orange juice case, the prohibitions on labeling and different bottles affect the importance of the price parameter. In the chocolate bar study, the different amount of level balance, especially for the filling attribute, causes large differences in the importance of these attributes.

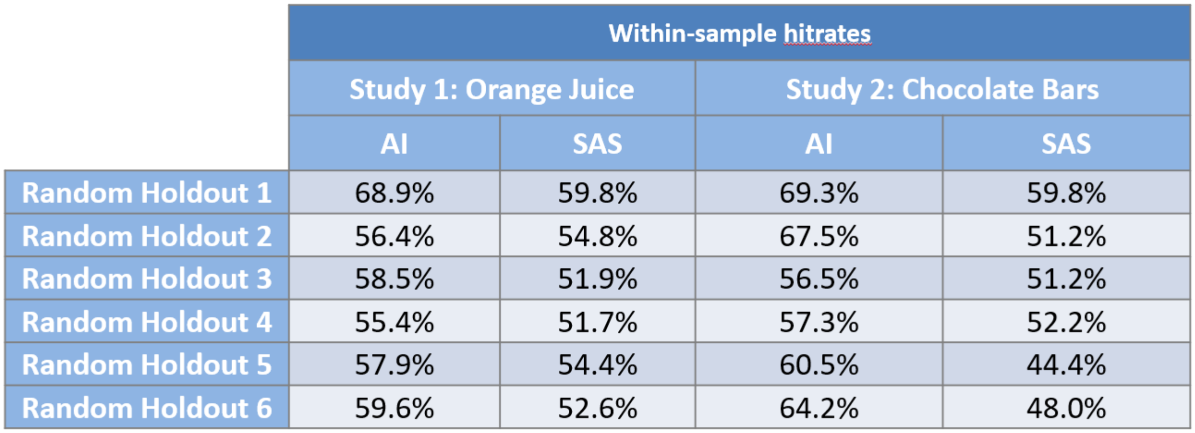
Figure 7: Attribute Importances from the Two Studies by the Two Design Strategies



Orange Juice Germany 2018/19; n=1010/1005; Chocolate bar 2018/2019; n=605/608

The comparison of in-sample hit rates is based on 6 selected random tasks (estimating the utilities 6 times by leaving one single random task out in each estimation). The result clearly showed that the hit rates on all holdout tasks are always better with the ANN-based versions (Table 5). The better handling of prohibitions and implausible combinations, which results in better level balance, seems to pay off.

Table 5: Comparison of Within-Sample Hit Rates by Study and Design

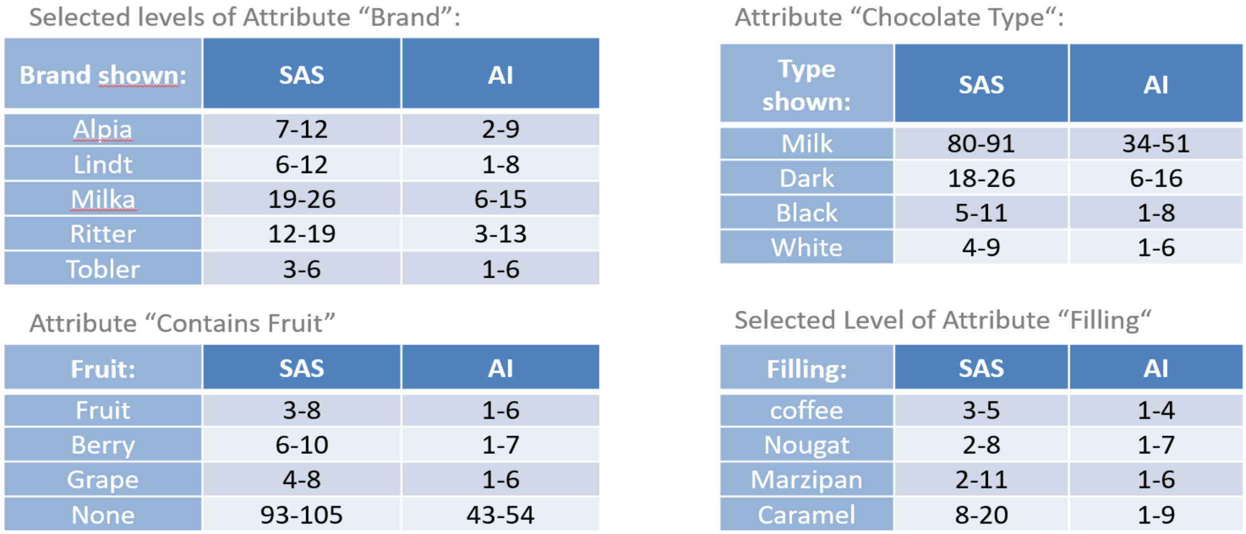


	Within-sample hitrates			
	Study 1: Orange Juice		Study 2: Chocolate Bars	
	AI	SAS	AI	SAS
Random Holdout 1	68.9%	59.8%	69.3%	59.8%
Random Holdout 2	56.4%	54.8%	67.5%	51.2%
Random Holdout 3	58.5%	51.9%	56.5%	51.2%
Random Holdout 4	55.4%	51.7%	57.3%	52.2%
Random Holdout 5	57.9%	54.4%	60.5%	44.4%
Random Holdout 6	59.6%	52.6%	64.2%	48.0%

Orange Juice Germany 2018/19; n=1010/1005; Chocolate bar 2018/2019; n=605/608

Looking deeper into level balance, we see that the deviation of one-way frequencies is larger for the SAS-based designs. Table 6 shows that the SAS-based designs always have a larger difference in the one-way frequencies of levels shown in one design version to the respondents. The superior level balance is one reason why AI designs are always a little better.

Table 6: Selected One-Way Frequencies from the Chocolate Bar Study



Selected levels of Attribute “Brand”:

Brand shown:	SAS	AI
Alpia	7-12	2-9
Lindt	6-12	1-8
Milka	19-26	6-15
Ritter	12-19	3-13
Tobler	3-6	1-6

Attribute “Chocolate Type“:

Type shown:	SAS	AI
Milk	80-91	34-51
Dark	18-26	6-16
Black	5-11	1-8
White	4-9	1-6

Attribute “Contains Fruit”

Fruit:	SAS	AI
Fruit	3-8	1-6
Berry	6-10	1-7
Grape	4-8	1-6
None	93-105	43-54

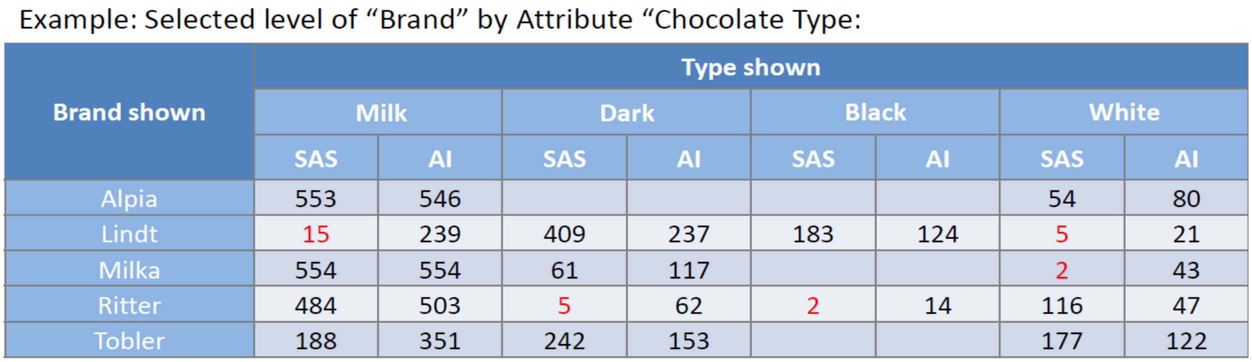
Selected Level of Attribute “Filling“

Filling:	SAS	AI
coffee	3-5	1-4
Nougat	2-8	1-7
Marzipan	2-11	1-6
Caramel	8-20	1-9

Chocolate bar study 2018/2019; n=605/608

Much more impressive are the differences in the two-way frequency tables (Table 7). For example, the attribute level “milk chocolate” is highly correlated with brands in the SAS design (only 15 views for Lindt, 554 for Milka). The AI-based designs do a much better job of achieving pairwise-level balance. Prohibitions of combinations didn’t affect the ANN design generation nearly as much as they affect the classical approaches.

Table 7: Two-Way Frequencies of Selected Attributes from the Chocolate Bar Study



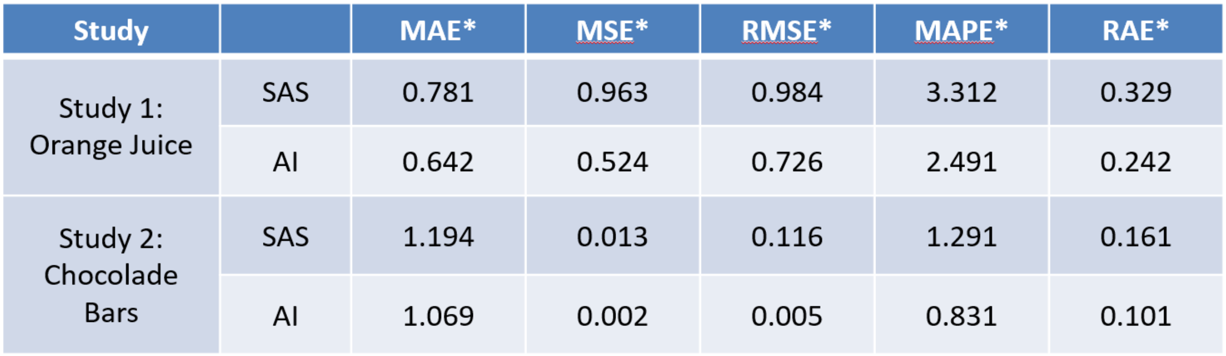
Example: Selected level of “Brand” by Attribute “Chocolate Type:

Brand shown	Type shown							
	Milk		Dark		Black		White	
	SAS	AI	SAS	AI	SAS	AI	SAS	AI
Alpia	553	546					54	80
Lindt	15	239	409	237	183	124	5	21
Milka	554	554	61	117			2	43
Ritter	484	503	5	62	2	14	116	47
Tobler	188	351	242	153			177	122

Chocolate bar study 2018/2019; n=605/608

For the chocolate bar study, we have real market data available and therefore we could do some out-of-sample predictions. We compared the base case share of choice for both designs with the market shares from the German chocolate market (Table 8).

Table 8: Error Measures[5] for Share of Choice for the Two Empirical Studies



Study		MAE*	MSE*	RMSE*	MAPE*	RAE*
Study 1: Orange Juice	SAS	0.781	0.963	0.984	3.312	0.329
	AI	0.642	0.524	0.726	2.491	0.242
Study 2: Chocolade Bars	SAS	1.194	0.013	0.116	1.291	0.161
	AI	1.069	0.002	0.005	0.831	0.101

Orange Juice Germany 2018/19; n=1010/1005; Chocolate bar 2018/2019; n=605/608

All error measures are slightly better for the results based on the AI designs. Measures like RMSE or MAPE that penalize larger deviations more than smaller ones especially show that the AI designs are superior to the classical ones. Both in-sample and out-of-sample results are better for the ANN-based design versions.

[5] All error measures are better when lower. MAE is mean absolute error. MSE is mean squared error (a variance-like measure that weights larger prediction errors more heavily than small ones). RMSE is root mean squared error, the square root of MSE and standard deviation-like. RMSE penalizes larger prediction errors more strongly but finally reports the average prediction error in the dimension of the original measurement units. MAPE is mean absolute percentage error (error as a percentage of the true value). RAE is relative absolute error, the ratio of an RMSE-like measure to the same measure for a naïve equal-probability model.

FINDINGS

As expected, the AI-based designs fit perfectly to the assumptions we included in our candidate generating process. In other words, the AI recovered the DGP. Prohibitions, implausible products, ordered attributes, assumptions about utilities, and pricing strategies were better accommodated better in the AI design versions. Overall, RMSEs, SDs, part-worth estimates, one-way and two-way frequencies, as well as D-efficiency are all closer to optimal designs in the AI-based approach than in the SAS-developed designs.

Adding a large amount of utility balance harmed the designs. Rich Johnson was right, we do not always gain an advantage by forcing utility balance in empirical data.

Small to medium amounts of utility balance (especially for ordinal attributes and price) results in better designs and showed no disadvantages on respondent burden and fatigue answering behavior.

From our two empirical studies we can conclude that there are differences in the estimated part-worth utility depending on the design generation technique. AI-based design techniques appear to deliver more stable and better results, although two empirical studies are not enough to conclude that they are always superior. We can see in our two examples that AI-based designs were superior in all tested measures. We see more valid results in case of attribute importance (proof of which is only possible in simulation studies) and more face validity, at least, for attribute importance in the empirical studies. Shares of choice (both simulated and in empirical studies) are closer to reality and have smaller errors. In our empirical studies we saw that the within-sample hit rates were higher when using AI-based choice tasks. In our second empirical study we saw that out-of-sample error between real market and predicted-share was slightly reduced with the ANN-based designs.[6]

In general, we can conclude that the Universal Approximation Theorem from ANN theory can be applied in the context of experimental design for Choice-Based Conjoint studies, meaning the ANNs are able to identify the DGP and come up with stable design versions. Although predictions generated by deep learning and many other AI technologies appear to be created from a black box, we can say that in our context the black box works well.

Most studies conducted in the marketing research community are still based on design algorithms which were developed for aggregate models. We have presented here a new technique which is able to generate optimized experimental designs for individual-level estimates.

FUTURE WORK

We need further investigations based on more empirical studies with out-of-sample data. We also need more work on what happens if we include wrong assumptions (such as too much utility balance)! So far, we can only suggest being careful when defining your assumptions for the candidates. We don’t really know what happens if the assumptions about utility balance are wrong.

[6] No data on real market share was available for the first empirical study.

A major potential step forward is the possibility of including "revealed preference" (past respondents' data) in the training of the ANNs to derive better designs for individual parameter estimates. If we have information from past choice models or panel data, we can train the second ANN with answers that have the same heterogeneity, same class memberships, and same utility structure as the revealed preference data. The trained ANN could then produce optimal deigns for fitting exactly to the actual respondent heterogeneity. ANN-generated designs based on past data could result in better input for pseudo-individual utility estimation (ability to capture more heterogeneity) by optimizing these designs for individual-level estimates.

If no past data are available, we can try, instead of using simple random answers, incorporating different answering routines in the computation, to come closer to real respondent answering behavior and the structure of real datasets, for use in the training phase of the ANNs.

As the proportion of None answers influences the results, we need further investigation of what happens if we assume a too large or too small None share in the training phase. Additional backpropagation loops to investigate the influence of None answers on the design during the training should be implemented and tested.

Peter Kurz

Stefan Binner

References

Agrawal, A.; Gans, J.; Goldfarb, A. (2018): *Prediction Machines*, Harvard Business Press, Boston.

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.

Alwosheel, A.; van Cranenburgh, S.; Chorus, C. (2017): Artificial neural networks as a means to accommodate decision rules in choice models, Presentation at the 5th International Choice Modelling Conference 2017.

Anderson, D.A.; Wiley, J.B. (1992): Efficient choice set designs for estimating cross-effects models, *Marketing Letters*, 3(4), 357–70.

Belyakov, D. (2018): Applying machine learning to CBC data, SKIM/Sawtooth Software European Conference Paris, April 2019.

Belyakov, D. (2017): kNNLogit Ensembles for CBC Studies, AMA ART Forum 2017.

Bishop, C. M. (1995): *Neural networks for pattern recognition*, Oxford University Press.

Bunch, D.S.; Louviere, J.J.; Anderson, D.A. (1994): A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes, working paper, Graduate school of Management, University of California at Davis.

Burgess, L.; Street, D.J. (2005): Optimal designs for choice experiments with asymmetric attributes, *Journal of Statistical Planning and Inference*, 134(1), 288–301.

Chorus, C.G. (2010): A new model of Random Regret Minimization, *European Journal of Transport and Infrastructure Research*, 10(2), 181–196.

Chorus, C.G. (2014): Capturing alternative decision rules in travel choice models: A critical discussion, Chapter 13 (pages 290–310) in Hess, S. & Daly, A. (Eds.) *Handbook of Choice Modelling*, Edward Elgar Pub.

Hensher, D. A.; Ton, T. T. (2000): A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice, *Transportation Research Part E*, 36(3), 155–172.

Hess, S.; Stathopoulos, A.; Daly, A. (2012): Allowing for heterogeneous decision rules in discrete choice models: an approach and four case studies, *Transportation*, 39(3), 565–591.

Hornik, K.,; Stinchcombe, M.; White, H. (1989): Multilayer feedforward networks are universal approximators, *Neural Networks*, 2(5), 359–366.

Huber, J.; Zwerina, K. (1996): The importance of utility balance and efficient choice designs, *Journal of Marketing Research*, 33(3), 307–17.

Johnson, R.M.; Huber, J.; Orme, B. (2005): A Second Test of Adaptive Choice-Based Conjoint Analysis (The Surprising Robustness of Standard CBC Designs). Proceedings of the 2004 Sawtooth Software Conference , Sawtooth Software Inc. Sequim WA.

Kuhfeld, W.F.; Tobias, R.D.; Garratt, M. (1994): Efficient experimental design with marketing research applications, *Journal of Marketing Research*, 43(3), 409–19.

Kuhfeld, W.F.: (2000): *Conjoint Analysis Examples*, SAS Institute TS-650F, Cary (NC).

Kuhfled, W.F. (1996): Marketing Research Methods in the SAS System, SAS Institute, Cary (NC).

Lazari, A.G.; Anderson, D.A. (1994): Design of Discrete Choice Experiments for Estimating Both Attribute and Availability Cross Effects, *Journal of Marketing Research* 31; 375–83.

Leong, W.; Hensher, D. A. (2012): Embedding decision heuristics in discrete choice models: A review, *Transport Reviews*, 32(3), 313–331.

Louviere, J.J.; Woodworth, G. (1983): Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data, *Journal of Marketing Research*, 20(4), 350–67.

McFadden, D. (2001): Economic choices, *The American Economic Review*, 91(3), 351–378.

Mohammadian, A.; Miller, E. (2002): Nested logit models and artificial neural networks for predicting household automobile choices: Comparison of performance, *Transportation Research Record*, (1807), 92–100.

Rao, Vithala R. (2014): Applied Conjoint Analysis. Springer Heidelberg, New York.

Teodorović, D.; Vukadinovic, K. (1998): *Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach*. Kluwer Academic Publishers, 1st edition.

van Cranenburgh, S.; Guevara, C. A.; Chorus, C. G. (2015): New insights on random regret minimization models, *Transportation Research Part A*, 74, 91–109.

Vapnik, V. N.; Chervonenkis, A. Y. (2015): On the uniform convergence of relative frequencies of events to their probabilities. In: Vovk,V.; Papadopolos,H.; Gammerman, A. (eds) *Measures of Complexity* (pp. 11–30), Springer, Cham.

PREDICTING THE (UNOBSERVED) PREDICTABLE: THE USE OF DEEP LEARNING IN WAVE STUDIES FOR MARKET RESEARCH

TOM GARDNER
MICHELLE MCNAMARA
ADELPHI RESEARCH

INTRODUCTION

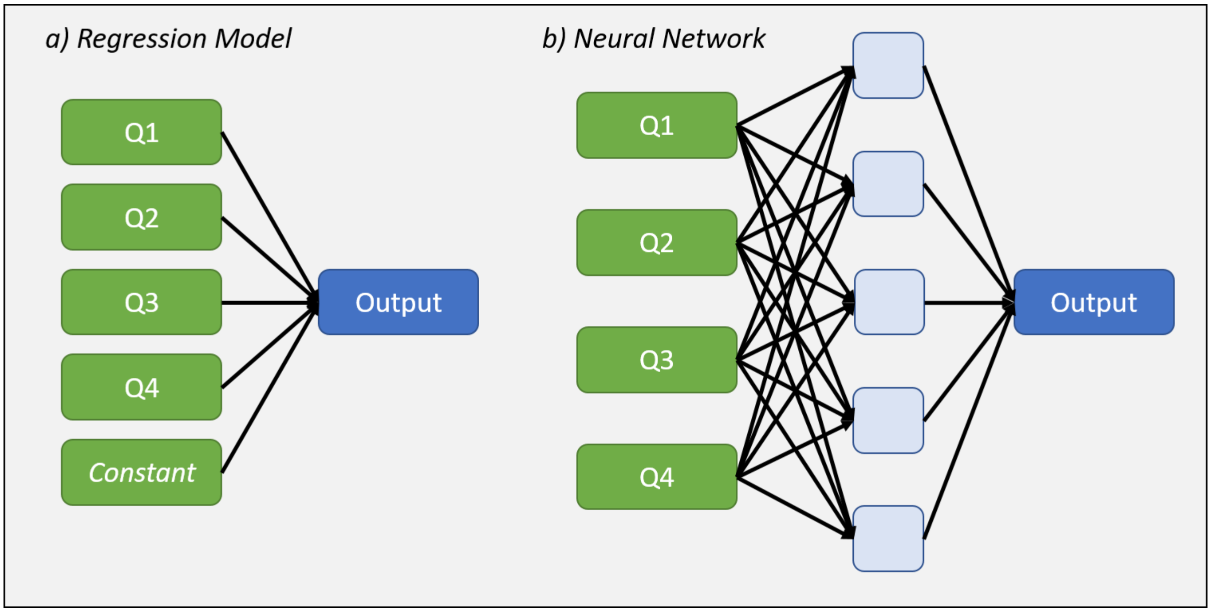
In Market Research, tracker (or wave) studies aim to understand the attitudes and usage of products over an extended period. The same questions are asked at the different time points, giving the client an understanding of how their brand performs over many months. With this dynamic information, clients can be informed on strategy and understand how their brand is situated within their competitive marketplace over a given time period.

One issue with this type of research is that it requires the recruitment of a large number of respondents. Pharmaceutical market research is an example of where recruiting large samples can be difficult and expensive. The reason for this is that the respondent pool is usually physicians, and their time is very expensive. Moreover, this sample pool is often further reduced by eligibility of respondents based on their specialty and practice setting. Nonetheless, clients still need to know how their products are performing over time so must engage in costly wave studies. From these relatively short online surveys, the client wants to extract as much information from the physicians as they can to make the most of their time and valuable opinions. The thesis of this manuscript is, given the cost and time constraints of physicians, there is a way to more fully optimise the benefits from each respondent.

To address this question, we thought about the nature of wave studies; they are repetitive and often there is little deviation for certain metrics over the time points. In other words, how physicians feel about the mode of administration of a drug is not likely to hugely fluctuate across the space of six months. In theory, if you could use the data obtained from previous waves to predict the behaviour in future waves, you need not ask the predictable questions in the latter waves, we can predict them from responses to the other attributes. This is an important concept because, if achievable, we could ask a subset of questions and get the data for a larger set—this would mean that we get more data from the expensive respondent interaction without asking more questions.

One method to address this paradigm would be to use a regression model. A regression model would be fitted to the data from waves 1 and 2 and then this model would be used to predict the omitted variable in wave 3. The issue with using a regression model in this way is that it will be looking for linear dependencies between the predictors and the dependent. A complicating factor for this kind of model is that time is factor; the relationship between the variables may change between the waves you are training the model on, and a regression model would not be able to model this nonlinearity in the data. As shown in Figure 1A, the input variables (in green) have a single coefficient applied to them to affect the output (dark blue), i.e., a linear dependency.

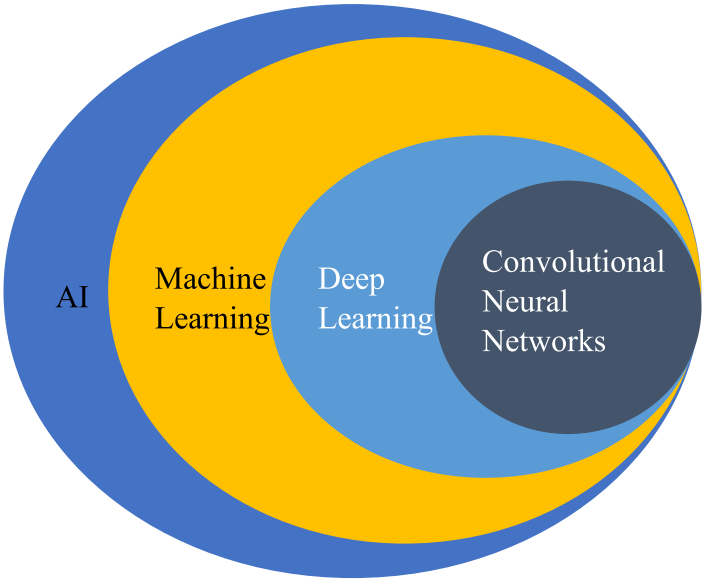
Figure 1: A) A schematic showing the relationship between predictors and output in a regression model. B) A schematic showing the architecture of a Neural Network.

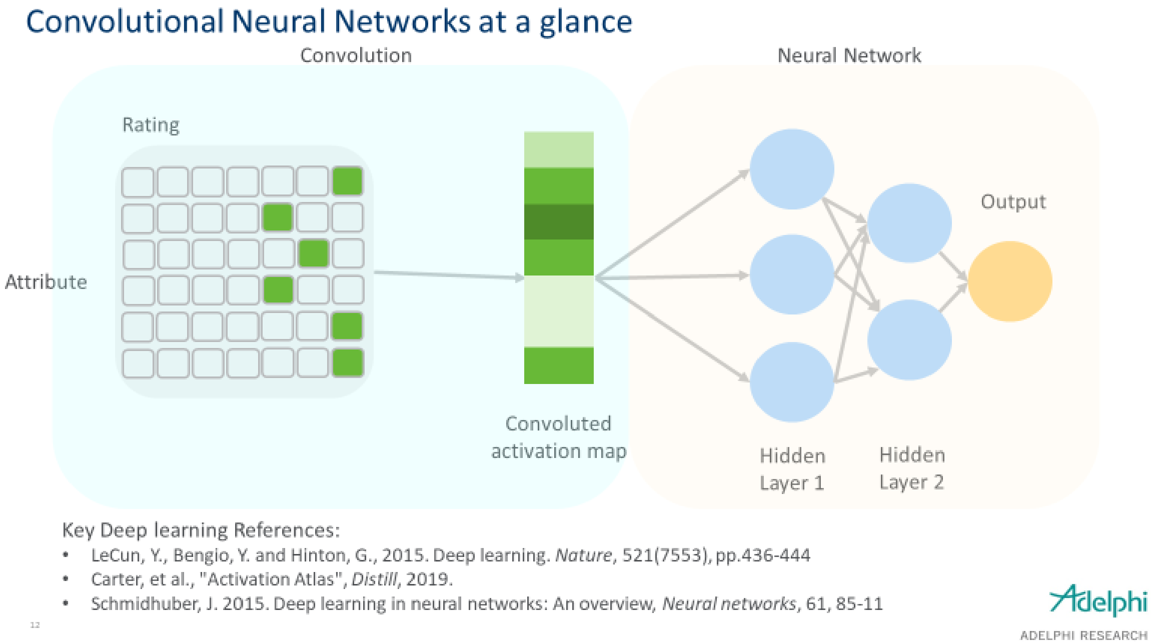


An alternative methodology is to apply a deep learning algorithm to this paradigm, namely fitting a Neural Network (NN). An NN can be seen as an extension of a logistic regression model as it allows for nonlinear and interactive functions (see Dreiseitl & Ohno-Machado, 2002, for a review and comparison of these methodologies). The "neural" aspect of an NN comes from our understanding of human brain function, namely that synaptic connections are nonlinear, and that input-output requires more than a direct connection between brain areas. For example, when a face is processed in the brain, information passes through several visual processing areas before we recognize the face as someone we know (see Rossion et al., 2012 for further explanation). A similar process is applied to NN outside of neuroscience, where to explain an output as a function of its inputs, multiple coefficients must be applied (or layers; light blue in Figure 1B), allowing nonlinear functions to be modelled (see Schmidhuber, 2015 for a review of Neural Networks). The power of this method comes from its ability to understand patterns which are not captured by traditional statistical techniques. Examples of these successful applications of NNs are recognising images (Krizhevsky et al., 2012), drug discovery (Ma et al., 2015), and natural language processing (Collobert et al., 2011).

Figure 2 shows a schematic of how Neural Networks fit into the field of Artificial Intelligence (AI). AI is an umbrella term applied to the use of machines for intelligent thought/making decisions. Deep learning and Neural Networks fall within the field of Machine Learning. In this study, we will use a form of Neural Network known as a Convolutional Neural Network (CNN). The way in which they differ from a standard NN can be seen in the bottom half of Figure 2, whereby the input into the model is different. It converts the raw input into an activation map, which is then passed into the model. This type of model is routinely used on tasks such as image classification and natural language processing to very high effect.

Figure 2: A schematic showing the field of AI (top) and a convolutional neural network (bottom).





Convolutional Neural Networks at a glance

Key Deep learning References:

- LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *Nature*, 521(7553), pp.436-444
- Carter, et al., "Activation Atlas", *Distill*, 2019.
- Schmidhuber, J. 2015. Deep learning in neural networks: An overview, *Neural networks*, 61, 85-11

12

METHOD

Data

The data used was taken from a long-standing wave study conducted by Adelphi Research. The purpose of this research was to find out the attitudes of physicians towards pharmaceuticals in Retinopathy. For this study, we have only used the respondents from the UK (see discussion). In each of the three waves there were 50 respondents (there was no overlap in this wave study, so each wave had unique respondents). We collapsed waves 1 and 2 together (known as the train data for the rest of this manuscript), and wave 3 (known as the test data) was set aside until the validation stages.

Variables and Identifying the Predictable Attributes

Each wave saw the same online survey, and therefore answered the exact same questions. In this survey, there was a battery of 17 attributes which respondents had to rate on a 1–7 Likert scale (Likert, 1932); these attributes were used to test our hypothesis. The task for respondents was to give a response between 1 (strongly disagree) and 7 (strongly agree) to the importance of each attribute in decision making when prescribing. The full list of these attributes can be found in Table 1.

Table 1: Battery of Variables Used

Variable Shorthand	Full text
Att 1	Long term efficacy
Att 2	Longest VEGF suppression
Att 3	Rapid visual gains
Att 4	Maintains a dry macula
Att 5	Most letters gained
Att 6	Proven safety profile
Att 7	Significant improvement in QoL
Att 8	Less frequent visits
Att 9	Flexible dosing regime
Att 10	Reduces injection burden
Att 11	Shorter injection preparation time
Att 12	Recommended in guidelines
Att 13	Leaders prefer the product
Att 14	Manufacturer provides patient support services
Att 15	Greater experience with the product
Att 16	Cost effective
Att 17	Delivers real world evidence
Att 18	Effective in a broad range of patients

From this list of 17 attributes, we wanted to identify a subset of attributes which could be predicted from the remaining attributes. Using the train data (waves 1 and 2), we conducted a Principle Components analysis (PCA; using a Varimax rotation) to identify components. The rationale here was that by identifying components and which attributes load onto them with the highest estimated correlation, we could omit these to be predicted as they serve as the archetypical variable for that component. The results from the PCA (Table 2) reveal three components which we have labelled efficacy, mode of administration, and credibility based on the variables which load to them.

Table 2: Reduction of Attributes by Using PCA (values indicate loadings, loading threshold of 0.1)

Attribute	Efficacy	Mode of administration	Credibility
Attribute 1		0.652	0.34
Attribute 2	0.481	0.432	0.411
Attribute 3	0.273	0.19	0.626
Attribute 4	0.677	0.195	0.41
Attribute 5	0.705		0.239
Attribute 6	0.796		0.215
Attribute 7	0.732	0.275	
Attribute 8	0.399	0.724	0.152
Attribute 9		0.78	
Attribute 10	0.532	0.582	
Attribute 11	0.128	0.649	0.27
Attribute 12		0.138	0.717
Attribute 13	0.311		0.657
Attribute 14			0.314
Attribute 15		0.272	0.533
Attribute 16	0.417	0.233	
Attribute 17	0.185	0.176	0.261
Attribute 18			

Highlighted in blue in Table 2 are the attributes for each component with the highest loading value, therefore those which we will aim to predict. However, upon reviewing the attributes which would be used, we made two changes. Firstly, we felt that “Less frequent visits” (Attribute 8) was semantically a consequence of Attribute 9 “Flexible dosing regime,” i.e., less visits is a consequence of a flexible dosing regime. As the estimated correlation for this attribute on the component was still very high, we replaced Attribute 9 with Attribute 8. Second, we decided to include Attribute 4 as one of the attributes to predict as this was an important attribute to our client. It had a strong estimated correlation with one of our components so we were confident that it could bootstrap to this component. Therefore, we took four attributes (highlighted in yellow) forward to predict.

Neural Network Parameters

The raw responses were first min/max standardised, removing any scale bias. The resulting values (between 0–1) were then converted into the convoluted activation map. This procedure used 3x3 kernels and the activation function Rectified Linear Unit (ReLU). This was conducted on the individual basis so that each respondent had a convoluted activation map which formed the input into the Neural Network.

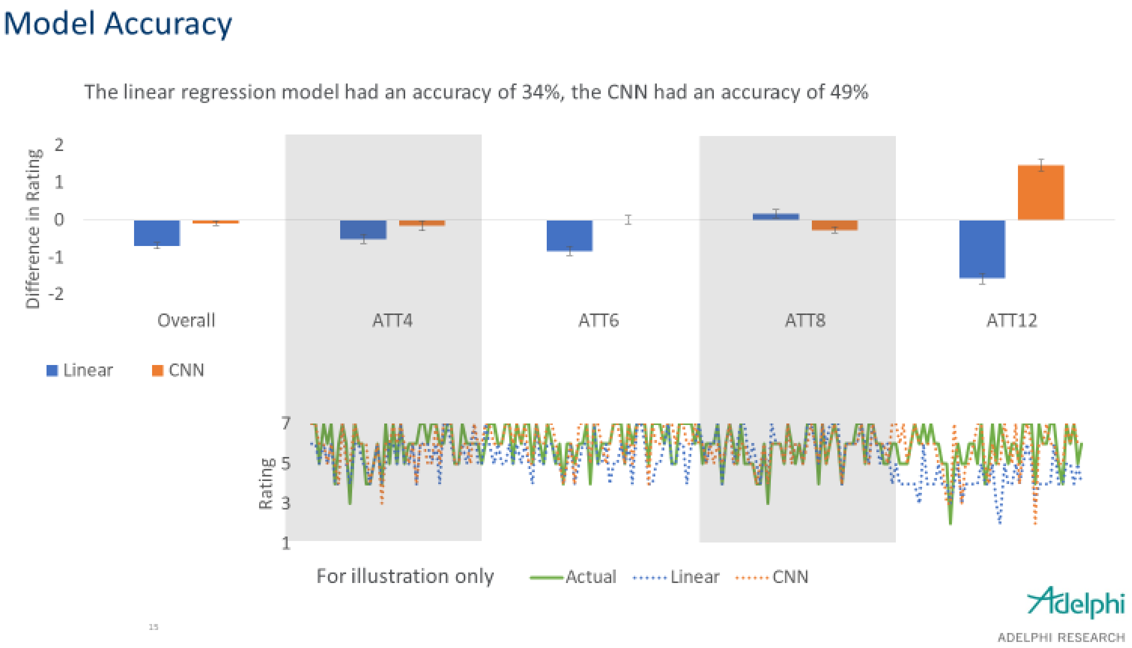
The CNN was created in Rstudio using the package NeuralNet (currently replicating using TensorFlow). The model architecture contained two hidden layers, with 30 and 15 neurons respectively. These parameters were determined by fine-tuning the model (checking overfitting of the model). The issue we faced was that we were using a small dataset which is not commonplace in the deep learning literature, however, it is a growing movement. The reason a small sample is problematic is that due to the complexity and unsupervised nature of deep learning algorithms, the model *will* fit the data, in other words, you can have a separate neuron for each respondent and fit the data perfectly. To avoid overfitting, we ensured the model had an accuracy of 75–85% in predicting the training data, across the 4 attributes we were predicting. This was checked by using the model to predict the data used to train the model.

Using the CNN model, we then tested our hypothesis by using our test data (wave 3) to see how accurately we could predict this future wave. As a reference point, we also used a linear regression model so we could see comparatively how well the CNN model performed.

Results

The respondent-level responses and predictions are shown in Figure 3. In the upper chart of this figure, we have plotted the difference in ratings between actual and predicted (positive is overestimated)—error bars are the Standard Error of the mean. From this chart we can see that the CNN appears to have predicted the future wave responses reasonably well. When we look at the hit rate of the CNN and linear model, the CNN has an accuracy level of 49% and the linear regression model has 34%. When we look at the raw data (lower chart of Figure 3), we can see that all models had issues predicting Attribute 12 accurately. The takeaway from this chart is that our model did a reasonable job; there is room for improvement though. However, clients do not report this respondent-level data, so it is less important to have an accurate prediction for that attribute.

Figure 3: The model accuracy. Upper chart shows the mean difference in rating for each of the attributes (for the linear regression model and CNN). The lower chart shows the raw responses plotted with the predictions of the models (for illustration purposes only).



Model Accuracy

The linear regression model had an accuracy of 34%, the CNN had an accuracy of 49%

For illustration only

When we look at the aggregate-level responses (Figure 4 and Table 3), we see that the CNN is incredibly accurate when predicting the actual responses. The reason we see this level of accuracy compared to the respondent-level accuracy is that the model has fit the data relatively well, however at the respondent level it would make mistakes, for example, predicting a rating of 6 when the actual response was 7. At the aggregate level, these inaccuracies are smoothed out and the aggregate-level accuracy is very precise. In terms of our hypothesis, this result would mean that we do not need ask about those 4 attributes, yet we can predict what the responses would be. If the client is reporting at the aggregate level, then this method would yield results on par with actually asking the respondent to rate the attribute.

Figure 4: The aggregate-level responses (mean) for the actual responses, CNN and linear regression model.

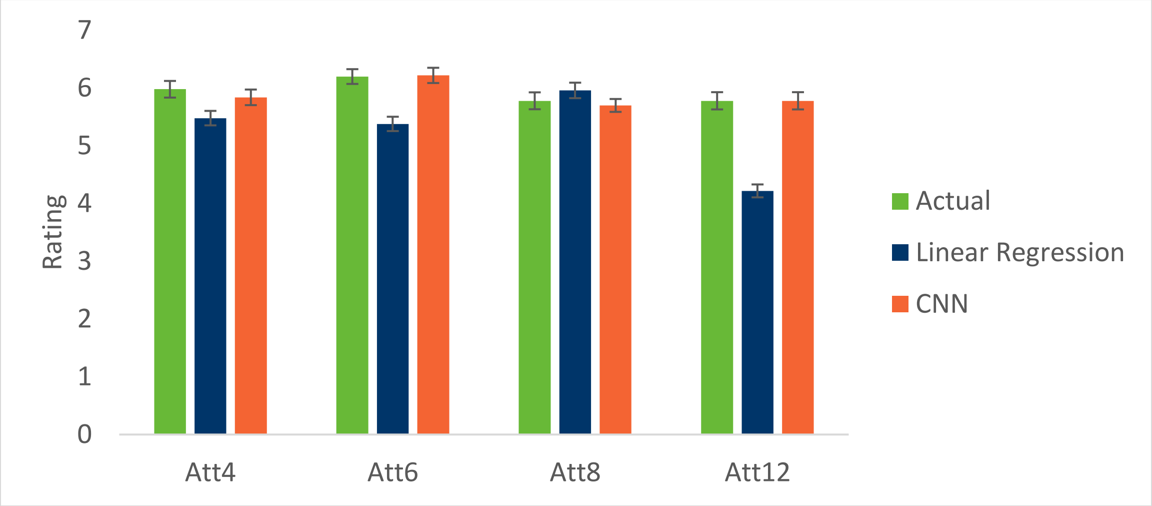


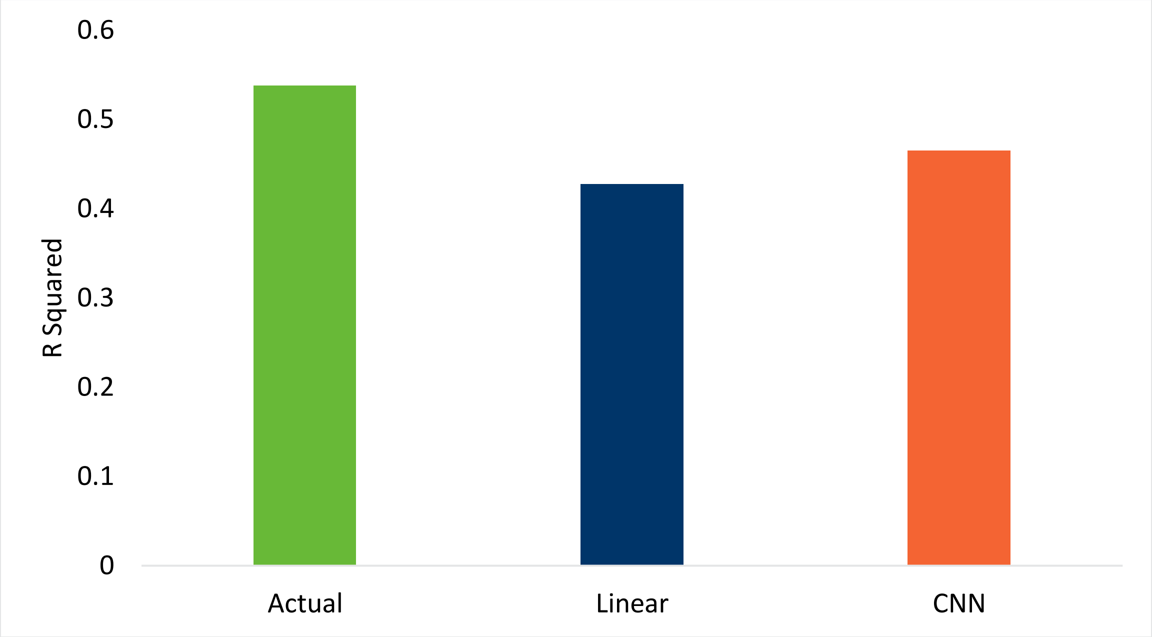
Table 3: Averages for each of the predicted attributes, for each of the models (including the actual).

Averages	Att 4	Att 6	Att 8	Att 12
Actual	5.98 (*SD*=1.02)	6.2 (*SD*=0.90)	5.78 (*SD*=1.04)	5.78 (*SD*=1.06)
Linear	5.48 (*SD*=0.89)	5.38 (*SD*=0.87)	5.96 (*SD*=0.95)	4.22 (*SD*=0.79)
CNN	5.84 (*SD*=0.96)	6.22 (*SD*=0.93)	5.7 (*SD*=0.79)	5.68 (*SD*=1.06)

Another aspect of these results to address is to confirm the appropriate use of deep learning and CNN; the reason we included a linear model was to address this point. The output shows that the CNN is highly accurate, however using a simple linear regression model would yield a less than desired level of accuracy. Furthermore, if the clients were to use aggregate-level responses to order these attributes based on rating, using the predictions from the linear regression model would give a very different story than the actual responses. Due to these findings, we can be confident in the use of a CNN model.

Finally, we wanted to test how well our model performed in explaining the variance in an external variable. By external we are referring to a variable which was not used in the creation of the CNN model. The reason we looked into this approach was the clients would often report these aggregate-level responses we have successfully predicted, but they also want to understand how these ratings relate to something like prescribing behaviour. A common approach here would be to use these ratings in a regression model to see how well they explain the prescribing behaviour of physicians. Using the respondent-level predictions from the CNN and linear models, we then tested the variance in prescribing behaviour explained (in terms of R Squared) of these compared to the actual responses. The results of this are shown in Figure 5.

Figure 5: The R squared values for each of our models when explaining prescribing behaviour.



The findings illustrated in Figure X show that the CNN model predictions are better than the predictions from a linear model. This improvement is moderate, however it should be noted that no prescribing related variables were present in the creation of these models, therefore these predictions are quite impressive. Furthermore, in an extension which will be discussed further below, we could include additional attributes in the creation of these models to better predict external variables if this is the aim of this approach.

Discussion

The aim of this study was to investigate whether there is a way to better utilize responses from physicians, given the high cost of each respondent. We leveraged a type of deep learning known as Convolutional Neural Networks to see if it was possible to omit attributes from a set of related judgments. In wave studies we were able to identify more predictable attributes, omit these, and then train a model to make predictions across waves. We tested CNN and linear regression on a holdout wave and found that the CNN more accurately predicted responses for the following wave. At an aggregate level CNN is very precise, and at the respondent level it helps explain the variance in an external variable.

There were certain issues in conducting this analysis. Firstly, it should be noted that when identifying the more predictable attributes which we would then go on to predict, PCA is not the only technique that could have been used to identify these variables. As this was an exploratory approach, we found this method helped identify the attributes that could be predicted. It provided a starting point that could be revised depending on managerial need for each attribute. Secondly, we believe we have taken sufficient steps in avoiding overfitting the data. We were acutely aware that overfitting the data could be an issue for CNN analysis, therefore by allowing the model to contain a certain level of error we ensured the model did not overfit. However, the downside of this method is that it is very time consuming. In a similar vein, creating an appropriate structure for the CNN model is also time consuming; this involves identifying the appropriate number of layers and neurons to allow the model to fit the data accurately but not overfit.

In addition to these issues, there are also some outstanding questions which need to be addressed. Firstly, and very simply, we acknowledge that reducing a battery of 17 attributes down to 13 saves relatively little time for our respondents. Thus, this current analysis provides an exploratory step to see if the number of attributes could be diminished. Greater accuracy could come from the addition of other variables such as physician specialty, years practicing, setting, and volume of patients seen. In addition, respondents also rate many different brands on batteries of attributes; this method could be used to reduce the responses needed by predicting the 4th brand from the other three.

Secondly, we are currently working on what happens if the respondents in the wave we are trying to predict don't behave how the others have before them? Or, what happens if an attribute suddenly becomes very important; can the model still predict correctly? The answer to this question is that no; no model can predict what it hasn't learnt. However, one way to limit this problem is by selecting variables that are strongly related to other variables. To the extent that their relationship remains constant, we should be able to continue to predict our target attributes. Still, it is the responsibility of ourselves and the client to be aware if there is an expected shift in the market, one where these attributes may no longer be predictable.

Conclusion

In conclusion, this method could be used to more efficiently collect data from a more expensive respondent resource, in our case physicians. We can use deep learning to predict ratings of attributes in future waves with less burden, with minimal impact on data quality. Such models can be used in primary market research to generate aggregate and respondent-level results. Surges in ratings between waves are hard to correct for but we are hopefully outlining attributes where this would not happen. Further work needs to be done on applying this to different data sets with different attribute lists, and a more comprehensive comparison between methods is needed. We have demonstrated that CNN is a viable solution but not the only solution.

Tom Gardner

Michelle McNamara

REFERENCES:

Carter, et al., “Activation Atlas,” *Distill*, 2019.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011). Natural Language Processing (Almost) from scratch. *Journal of Machine Learning Research*, 12, 2493–2537.

Dreiseitl, S. & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodological review. *J Biomed Inform*, 35(5–6), 352–9.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems.* 25. 10.1145/3065386.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning, *Nature*, vol. 521, no. 7553, pp. 436–444.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140, 1–55.

Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model*, 55(2), 263–74.

Rossion, B., Hanseeuw, B., & Dricot, L. (2012). Defining face perception areas in the human brain: A large scale factorial fMRI face localizer analysis. *Brain and Cognition*. 79 (2): 138–157.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview, *Neural networks*, 61, 85–11.

CAN WE REDUCE THE NUMBER OF TASKS AND STILL GET GOOD QUALITY RESULTS?

CHRIS MOORE
IOANNIS TSALAMANIS
IPSOS MORI UK

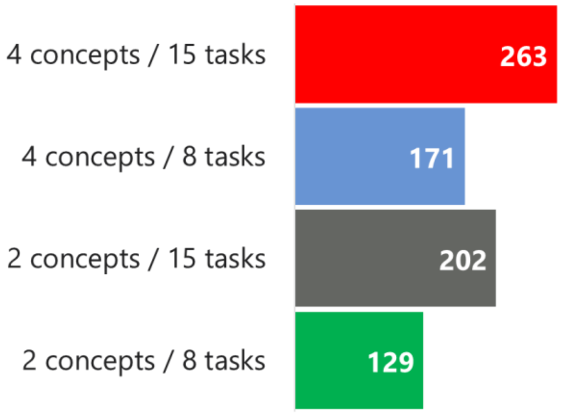
BACKGROUND

Within the market research industry there has been an increasing shift towards projects being delivered quicker and more cost efficiently, while still obtaining the same high standards. This, combined with respondents making more use of mobile devices to answer surveys, has led to the need to reduce questionnaire length. The long-held belief that you can ask respondents to complete 20-minute online interviews has never been more challenged and a new paradigm shift to surveys of c.10 minutes is more likely to become the norm, rather than the exception in the future. While the time needed to conduct a conjoint exercise can vary considerably, based on the type of conjoint being used, and the size of the design, it is not uncommon for a Choice-Based Conjoint (CBC) to take 5 or more minutes to conduct, taking up at least half of the questionnaire time. As a result, there is a need to investigate how we can reduce the time needed to conduct a conjoint exercise or else we run into the danger that conjoint becomes a nice-to-have rather than an integral part of the research.

It is now becoming common practice to create modular questionnaire designs, where the data for modules of questions that are not asked to a respondent is imputed. Therefore, the question is whether we can apply the same thinking to conjoint designs and still get similar results? Hierarchical Bayes (HB) analysis has consistently shown to be durable with sparse designs; but with sophisticated imputation procedures commonly available in R packages or Python libraries, it is hypothesized that creating additional respondent data (as inputs to a HB estimation) using these imputation procedures, which will benefit from being able to use many of the variables from the rest of the survey as predictors in the imputation process, will result in only a minimal loss of accuracy compared to a task-rich design.

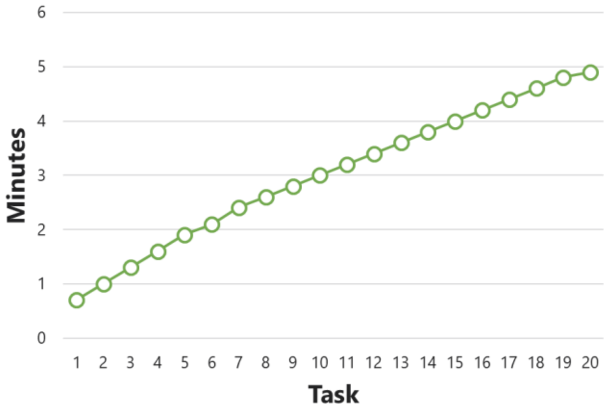
From internal data, looking across major markets such as the US, UK, and France, almost two-thirds of interviews conducted in 2019 were less than 15 minutes in length, with a mean time of 11 minutes. This is a far cry from historic figures where it was considered that 20 minutes for an online interview was the acceptable norm. By contrast, when we look at the time to complete a conjoint survey we see a different story. A study that was reported at the Sawtooth Software conference in 2016 (Moore/Neuerburg) was conducted with 6,800 online interviews, where respondents were split across 18 experimental cells, answering primarily on mobile devices (mobile/tablet), and the time taken in seconds to complete the exercise for different experimental cells is shown in Figure 1.

Figure 1



It should be noted that half of the respondents answered the survey using an angular JS responsive technology, that significantly increased the speed of the conjoint exercise as it is a single-page application, which does not require communication with the server after each click. For respondents that went through the more traditional Dimensions (SPSS) platform, across all experimental cells, it took on average 31 seconds extra to complete the exercise. Therefore, a conjoint exercise in the region of 5 minutes is not unexpected for a 4 concept/15 task design. Johnson/Orme (1996) published a graph of the time taken by respondents to complete a 20-task conjoint exercise, containing 3 concepts plus None option (Figure 2), and when comparing against the results from the Moore/Neuerburg study, it is reasonable to infer that the time taken for an 8 or 15 task study is very comparable.

Figure 2



While we know that length of interviews in the industry are declining, despite respondents increasingly moving into a digital environment and using mobile devices to answer surveys, the length of time to complete conjoint exercises has not decreased. Therefore, combined with trying to reduce respondent fatigue we need to ask ourselves the question of whether we can reduce the time needed to conduct a conjoint and still get high quality results?

STUDY DESIGN

Five existing data sets were identified for this research, each of which had specific criteria: that there were a low number of conjoint versions (blocks) tested within the design and that there were sufficiently high numbers of respondents that had answered each version. This is because for the imputation process to work it is a requirement that you can only impute missing task data for a respondent based on other respondents that have answered the same version of tasks. The larger the number of respondents that have seen each version, the higher the likelihood of a successful imputation. Several different experimental factors were tested in this research (Figure 3). These included:

Methods Tested:

The two main methods evaluated were that of a sparse design and an imputed design. The sparse design differs from a typical sparse design, in that while the number of tasks has been reduced, respondents were randomly allocated a set of tasks from the larger pool of tasks; e.g., across a full 9 tasks design respondents randomly had data for 5 tasks removed, but not the last 5 tasks. There are pros and cons to this approach:

- Respondents typically change their behaviour throughout a choice experiment so by not taking the first X tasks (as per a standard sparse design), we are including task data which may have been answered differently compared to earlier tasks.
- However, if we only take the first X tasks then we have no available data in the tasks that were not asked about and therefore it would be impossible to conduct any imputation.
- While understanding the implication of the first point it was felt that the approach used offered a more like-for-like comparison between the sparse and imputed method because the same valid task data was being used in both methods.

For those tasks that had been removed in the sparse design, they went into the CBC/HB software as missing data. For the imputed designs, where a respondent had not answered a task from the pool of tasks in the version they were assigned to, imputation procedures were used to predict what the respondent would have answered, based on responses to other variables in the survey. As such, a larger number of tasks were used for the CBC/HB estimation than for the sparse design. Further details of how the sparse data set was created are shown later in the paper.

As a second dimension, during the CBC/HB estimation process, the analysis was run both with and without covariates for both the sparse and imputed designs. Covariates were selected based on the known results when estimating the conjoint with the full and complete set of tasks and understanding which groups of respondents showed the most discrimination in results.

Imputation Method:

For the purposes of this study, two imputation procedures were initially tested. One is a model-based imputation that creates a regression model to identify nearest neighbours based on the predicted value. The other is a distance-based method, that identifies the distances between respondents based on subjective decisions about how respondents have answered other survey questions.

Number of Tasks Removed:

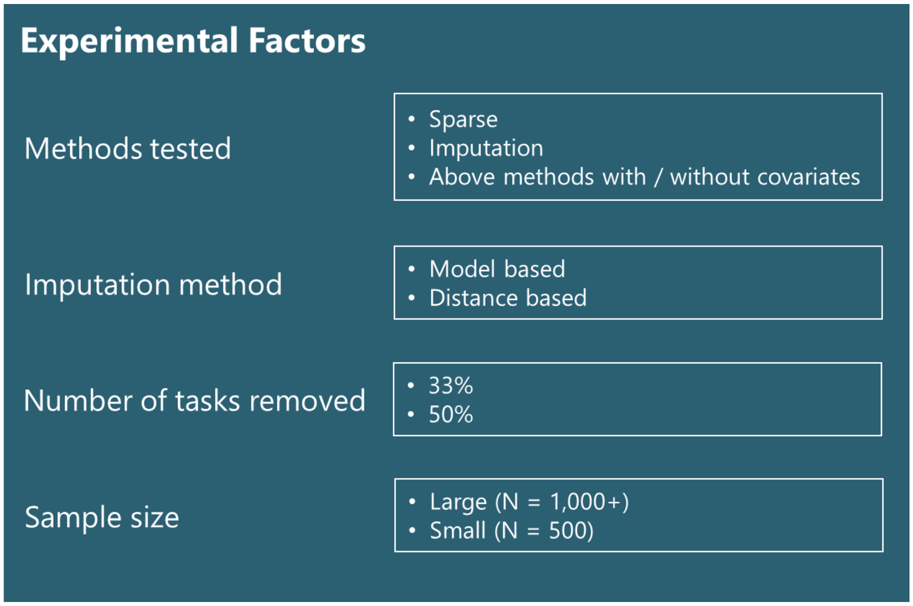
The original design for these studies either had half of the tasks removed or a third of the tasks removed. Where the original design did not allow for a perfect split, e.g., removing 50% of tasks from a 9-task study, the number was rounded down, so in this example of 9 tasks, 5 tasks were removed.

Sample Size:

The data sets identified were already rich in sample (greater than N = 1,000), so in addition to running analysis on these data sets, a random sample of N = 500 respondents were taken from each data set to generate a further 5 data sets.

Therefore, 8 sets of analysis were run for each of the 10 data sets, in addition to the original analysis containing the full and complete data set.

Figure 3



Experimental Factors

Methods tested	• Sparse • Imputation • Above methods with / without covariates
Imputation method	• Model based • Distance based
Number of tasks removed	• 33% • 50%
Sample size	• Large (N = 1,000+) • Small (N = 500)

How Does Imputation work?

Imputation and fusion are often confused with one another, and for very good reason as they have a very similar methodological background. For a typical data fusion project, two data sets would have a common set of variables X, while data set 1 also contains a set of variables Y and data set 2 contains a set of variables Z. The goal of fusion is to create a single data set that contains the X, Y, and Z variables, such that the correlation/relationship between Y and Z can be observed, where previously it could not.

Imputation, on the other hand, is typically used for data enrichment. In this case, there is a data set that contains a set of variables X and a set of variables Y, and another data set, or a set of respondents from the same data set, that have a complete set of X variables but no information about the Y variables. The goal is to therefore create a single data set where each observation has a complete set of X and Y variables, whilst preserving the correlation/relationship structure of X and Y. It is often referred to as "Assisted" fusion, as the correlation between X and Y is already known, whereas in data fusion, the correlation between Y and Z is unknown.

There are many imputation methods that exist, which can be easily accessed via software such as R and Python. Popular R packages include mice, Amelia, BaBooN, mi, mitools, pan, and missForest.

Model-Based Imputation:

Depending on the package used, the method for imputing will vary but a typical process would involve:

1. Replace all missing data for a variable with random draws from the observed data for that variable.
2. For all variables with missing data, build a (regression) model using other variables in the data set as predictors.
3. Where there is missing data, identify similar respondent(s) based on the predicted values from the model.
4. Substitute any missing data with observations from the similar respondent(s).

Steps 2–4 can then be repeated if necessary.

Distance-Based Imputation:

Distance-based methods are more synonymous with data fusion techniques and are based on finding similarities between respondents and using these similarity measures as proxies for filling in missing responses. The similarities between respondents can be measured between different groups of variables depending on the study. These groups of variables can then be weighted in a way that the analyst believes is appropriate for the sample, or a weighting scheme can be chosen that provides the maximum efficiency, which in this case is classifying someone to the correct concept chosen.

For this study, the variables were divided in two distinct groups and a similarity matrix was generated for each group. The first group of variables consisted of all the demographic questions such as age and gender, while the second group consisted of all the responses from the rest of the study. For categorical data, respondents were given a score of 1 if they matched on a variable and 0 if not, while for numeric/ordinal data, a score between 0–1 is calculated depending on how close the responses are. The scores are then added up to create the similarity matrix between all respondents and the matrix data is normalized by dividing by the mean score in the sample. Once the matrix for each group of questions is derived, the matrices are then weighted and added together to provide the final matrix that will be used to find the best matches. A grid search approach was used to identify the best weighted combination that provided the best accuracy measure. From the analysis, the ratio of the weighting between the demographic and covariate matrices was 1:6.

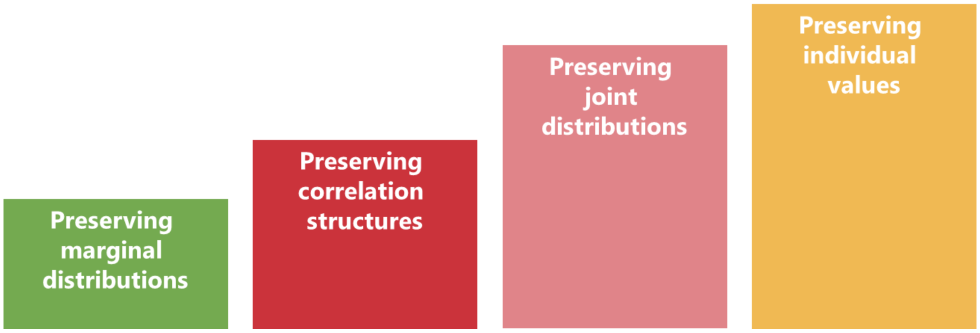
From the tests conducted, the model-based approach provided superior accuracy and results in this research are based on that method.

Initial Thoughts Going in to the Research

Imputation has been growing in popularity over several years and won an MRS award in the UK in 2017 for the use of imputation in reducing the length of surveys by up to a third. Internal analysis has shown that it works very well when data conditions allow. That is, when variables to be imputed are numerical/ordinal and a plethora of relevant data is available to be used as explanatory variables.

Regarding the success of any imputation procedure, Figure 4 (Rassler, 2002) shows the typical steps used to measure success. As an entry point, preserving the marginal distribution of the variable is a key factor, followed by preserving the correlation structure across variables. After that, preserving the joint distributions and individual scores can also be used as further criteria.

Figure 4


Preserving marginal distributions
Preserving correlation structures
Preserving joint distributions
Preserving individual values

Going in to this research there were four known limitations:

1. Choice data is categorical so would be harder to predict.
2. Instead of being able to use all respondents, it is only possible to impute data from respondents that answered the same versions of tasks.
3. The survey questionnaire had not been set up specifically with imputation in mind and may not have enough variables of good quality to use for the imputation.
4. The success criterion of the imputation in this research is solely on the ability to accurately classify a respondent with the correct concept.

While there were several factors that might suggest imputation is not appropriate, there were a few other considerations:

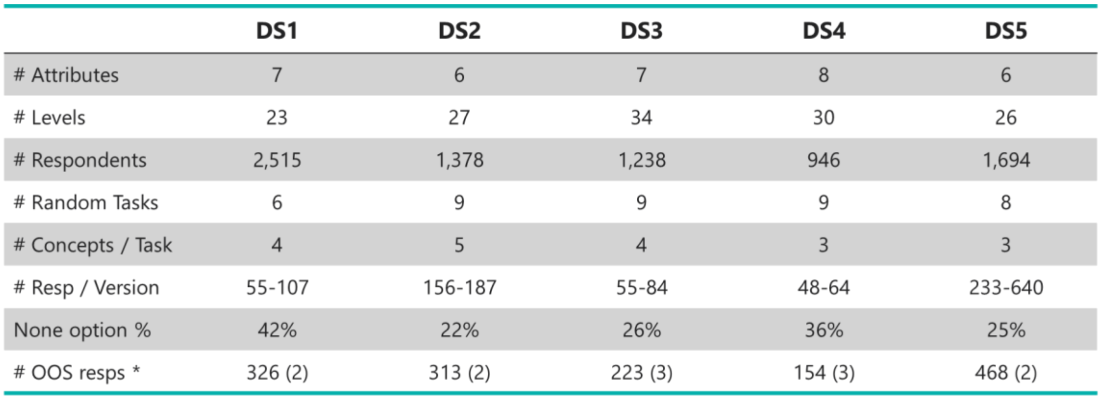
1. After cleaning the data, there were between 30-100 variables that could be used to build the explanatory model.

2. HB acts as an imputation of sorts, in that it adjusts the parameters of the individual respondent based on the upper-level model. As such, it was anticipated that it would smooth out issues with the imputed data where respondents had been incorrectly classified with the wrong concept chosen.

Data Sets

A summary of the data sets used is shown in Figure 5. All data sets were based on a CBC design methodology and were similar in terms of complexity. None of the data sets included advanced design features such as alternative-specific designs, and attribute levels were analysed as part-worths using effects coding (a variation on dummy coding).

Figure 5



	DS1	DS2	DS3	DS4	DS5
# Attributes	7	6	7	8	6
# Levels	23	27	34	30	26
# Respondents	2,515	1,378	1,238	946	1,694
# Random Tasks	6	9	9	9	8
# Concepts / Task	4	5	4	3	3
# Resp / Version	55-107	156-187	55-84	48-64	233-640
None option %	42%	22%	26%	36%	25%
# OOS resps *	326 (2)	313 (2)	223 (3)	154 (3)	468 (2)

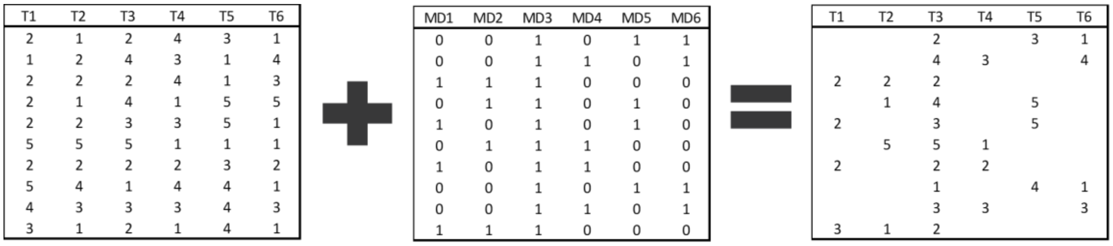
Other than one version, in one of the data sets, for all versions of tasks there were at least 50 respondents that evaluated each version, and in two of the data sets there were more than 150 respondents evaluating each version. An out-of-sample (OOS) data set was generated from holding out respondents from 2-3 versions of the original design at the analysis stage.

Data Preparation

To create the sparse designs, a MaxDiff design was created to ensure that across the sample, each combination of tasks had responses to them and to reduce any bias. It was hypothesized that data from other tasks would have a positive influence in the imputation process so other than data set 1 (1 task), two tasks were left untouched and therefore had full data available. An example of the process is shown in Figure 6.

The resulting file was then used for the sparse analysis, where data that had been removed went in as missing data in the HB estimation. For the imputation estimation, the sparse design was fed into R software, along with data from the rest of the survey, and the imputation procedure was applied.

Figure 6



T1	T2	T3	T4	T5	T6
2	1	2	4	3	1
1	2	4	3	1	4
2	2	2	4	1	3
2	1	4	1	5	5
2	2	3	3	5	1
5	5	5	1	1	1
2	2	2	2	3	2
5	4	1	4	4	1
4	3	3	3	4	3
3	1	2	1	4	1

+

MD1	MD2	MD3	MD4	MD5	MD6
0	0	1	0	1	1
0	0	1	1	0	1
1	1	1	0	0	0
0	1	1	0	1	0
1	0	1	0	1	0
0	1	1	1	0	0
1	0	1	1	0	0
0	0	1	0	1	1
0	0	1	1	0	1
1	1	1	0	0	0

=

T1	T2	T3	T4	T5	T6
		2		3	1
		4	3		4
2	2	2			
	1	4		5	
2		3		5	
	5	5	1		
2		2	2		
		1		4	1
		3	3		3
3	1	2			

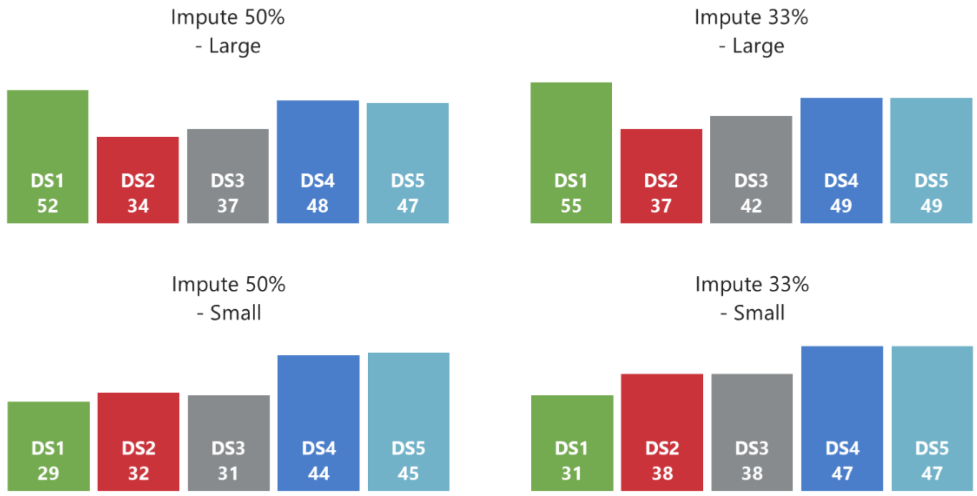
IMPUTATION RESULTS

Checks of the marginal distributions (percentage of respondents that selected each concept for each task in the estimation) pre and post imputation showed that the model-based imputation had done an excellent job of preserving the original distributions. For the large sample sizes, where 50% of the tasks had been removed, the error rate across the 5 data sets was between 0.7%–1.8%. Even in the sparsest conditions (small data set, 50% tasks removed), the error rate only ranged between 1.5%–3.2%.

While it is important to understand that the imputation procedure has done an excellent job of preserving the distribution of responses, the ultimate success of this research is a facet of the hit rate, which is the proportion of times we correctly classified missing data with the correct concept.

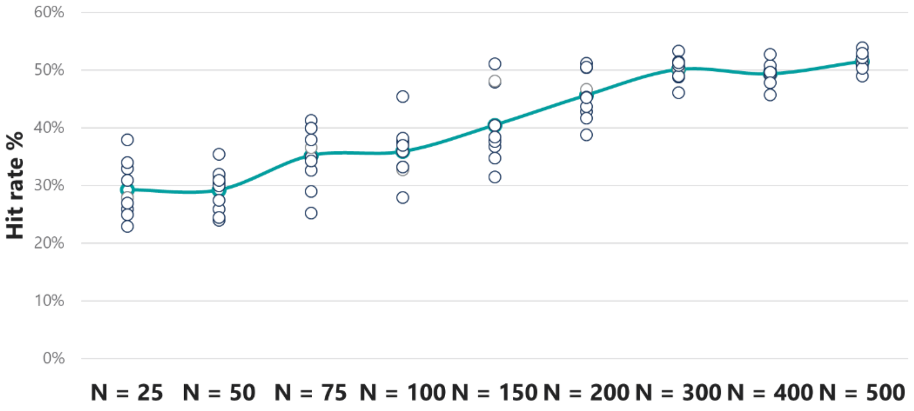
Figure 7 shows the hit rate for each of the 10 data sets used in the study. On average, for the large data sets, the imputation analysis resulted in correctly classifying data approximately twice as well as chance (note: data sets contained between 3-5 concepts plus the None option). Data set 1 had the best classification rate with 2.5 times better than chance. It is suspected that this may be due to the high proportion of tasks that were answered with the None option. The results across the data sets where we removed 33% and 50% of the data were very similar, indicating that the presence of additional tasks only had a marginal effect. When looking at the imputation success of the smaller data sets, the results are very aligned to the respective larger data set, other than in data set 1, where the imputation hit rate was extremely poor. It was not clear after carrying out an investigation as to the cause of this, but this data set had the sparsest conditions (6 tasks in the original design), and the largest difference in sample size between the two data sets (original sample size was N = 2,515), so a number of the important correlations in the original data set may have been impacted.

Figure 7



It was unexpected to see that the hit rate for the smaller data sets (except data set 1) were only marginally lower than the respective larger data set, with most differences being less than 4 percentage points. To understand this result further, using a design version from data set 5, which had over 500 respondents, 10 sets of imputations were run for each of 10 different sample sizes, ranging between N=25 and N=500. For each replication, respondents were randomly drawn without replacement. The resulting hit rate was then recorded (circles in Figure 8). As expected, as sample size increases then the average hit rate improved, though it is noticeable that the range of hit rates at the lower sample sizes varied significantly. In some instances, where the sample was N = 50 or less, results were worse than chance. The range of hit rates obtained stabilised from N= 200 onwards, though it is interesting to note that once the sample size went above N = 300, the hit rate did not improve significantly.

Figure 8



RESULTS

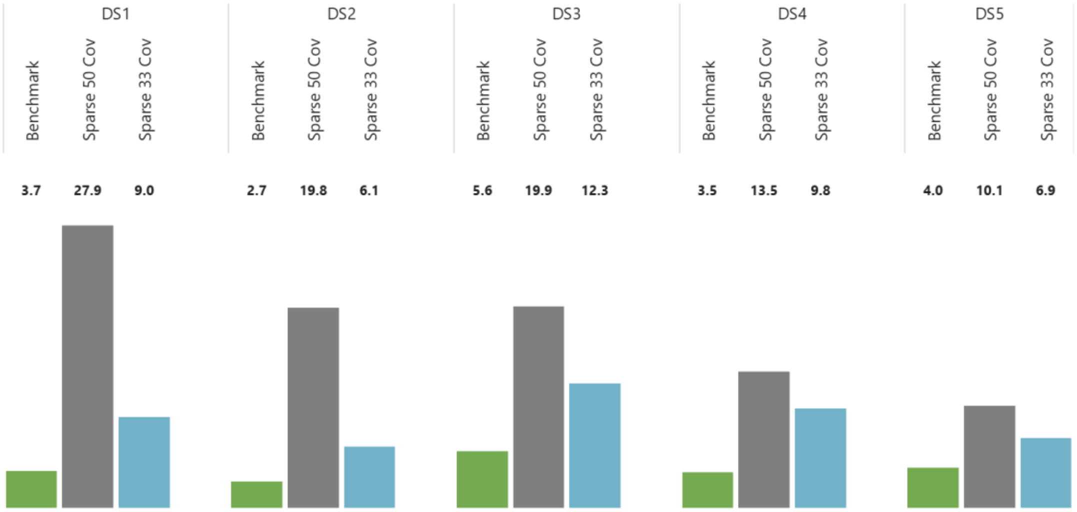
After running the analysis through CBC/HB, several diagnostics were compared. One of these metrics is Percent Certainty. This figure indicates how much better the solution is than chance, as compared to a "perfect" solution. It is equal to the difference between the final log likelihood and the log likelihood of a chance model, divided by the negative of the log likelihood for a chance model. It varies between zero and one, with a value of zero meaning that the model fits the data at only the chance level, and a value of one meaning perfect fit (Sawtooth Software CBC/HB manual).

The results of each model were compared against the benchmark, which was the original data containing the full set of tasks. It is only possible to compare the Percent Certainty figures for models that have the same number of tasks used in the estimation procedure, so only the imputed cells were compared against this benchmark data set. Across the 5 large data sets, there was a reduction of 21% in fit for the cells where 50% of the tasks were imputed. This compared to a reduction of 14% in fit for cells where 33% of the tasks were imputed. Adding in covariates increased the fit of the models marginally but little should be made of this as Sentis and Geller (2010) showed that it is possible to create random data and enter this data as covariates, and it will increase the fit. After running tests on several data sets, using random data as covariates, the same conclusion was also found in this research.

With a logit model the scaling of the part-worth utilities depends on the goodness of fit: the better the fit, the larger the estimated parameters. Thus, the absolute magnitude of the parameter estimates can be used as an indicator of fit. Average Variance is the average of the variances of part-worth utilities, across all respondents (Sawtooth Software CBC/HB manual). Due to the error from incorrectly specifying the correct concept chosen from the imputation process it was expected that the average variance for the imputed cells will be lower than the benchmark and sparse designs. While expecting to see a reduction, the actual variance recorded was significantly lower. Even in conditions more conducive to imputation (data set 5), across the 4 imputed cells, there was a 35% reduction in variance on average. In the most extreme case (data set 3) a reduction of 62% in variance was recorded.

For the large data sets, the variance in cells where covariates were present did not significantly impact the scaling of the parameters. This was not the case for the small data sets, where the addition of covariates led to a significant increase in the size of the part-worth parameters. Figure 9 shows the average variance for each of the conditions by data set, where the average variance of the parameters, particularly in the case where 50% of tasks were removed, led to 4- to 6-fold increase in variance, suggesting that there is significant overfitting occurring.

Figure 9



While it is important to understand changes in goodness of fit regarding sparse and imputed data sets, from a business aspect it is imperative that the business outcome is the same. To test this, the individual respondent-level part-worth utilities for each experimental cell were stacked into a single column of data and correlated against the respondent-level part-worth utilities from the benchmark data set. It was surprising to see that for the sparse 50 experimental cell that the correlations were very high, with values ranging between 0.84-0.94, across the 5 large data sets. This compared very favourably with the imputed 50 experimental cell, which recorded correlations between 0.67-0.88. However, in all large data sets the clear winner was the sparse 33 experimental cell where the correlation varied between 0.93-0.96. As expected, the correlations in the cells where covariates were applied were much lower due to the nature of how covariates work in CBC/HB, and that the upper model is different for different segments of the sample.

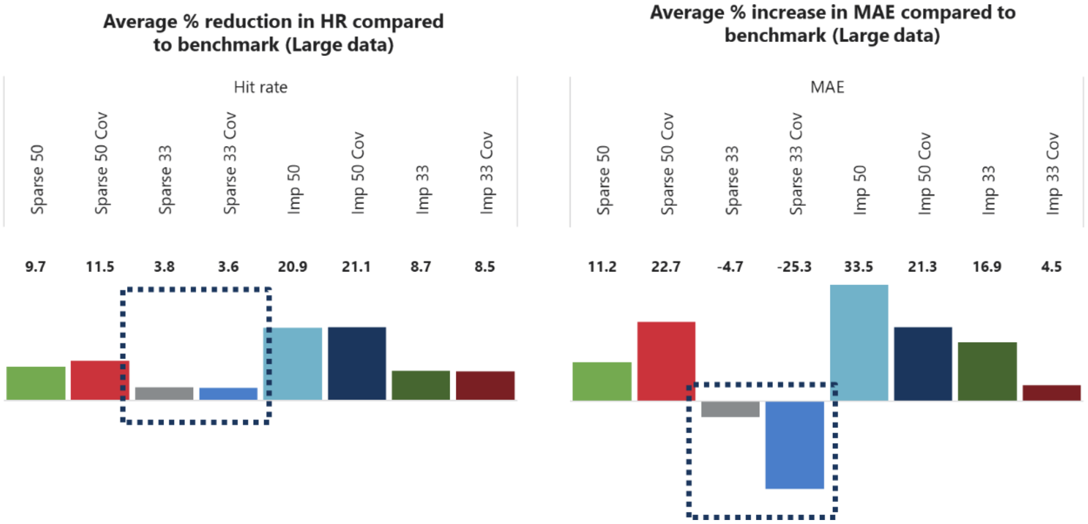
Four of the data sets contained at least one holdout task, which could be used to review changes in hit rate and Mean Absolute Error (MAE). While it is more desirable and robust to include 5-6 holdout tasks (compare to the 1-2 in these data sets), as these were commercial data sets it was not feasible to include that many holdout tasks. However, the results are similar across the data sets, so it is possible to still obtain valuable insight from the results.

Figure 10 shows the differences in hit rate and MAE compared to the benchmark data set. For hit rate, it is measured as the average percentage reduction in hit rate compared to the benchmark (large data sets) and for MAE it measured the percentage increase in MAE compared to the benchmark. For the MAE analysis, the data was adjusted for scale. As discussed earlier, due to the additional response error in the imputed data sets, the variance of the part-worth parameters is significantly lower and therefore will impact MAE. To allow comparisons across models, an exponent factor is included in the share of preference calculation to minimise the MAE score.

The sparse 33 experimental cell is the clear winner in both measures with a reduction of only 3.6-3.8% in hit rate (depending on if covariates were included or not). This is followed by the imputed 33 experimental cell which saw a reduction of 8.5-8.7%.

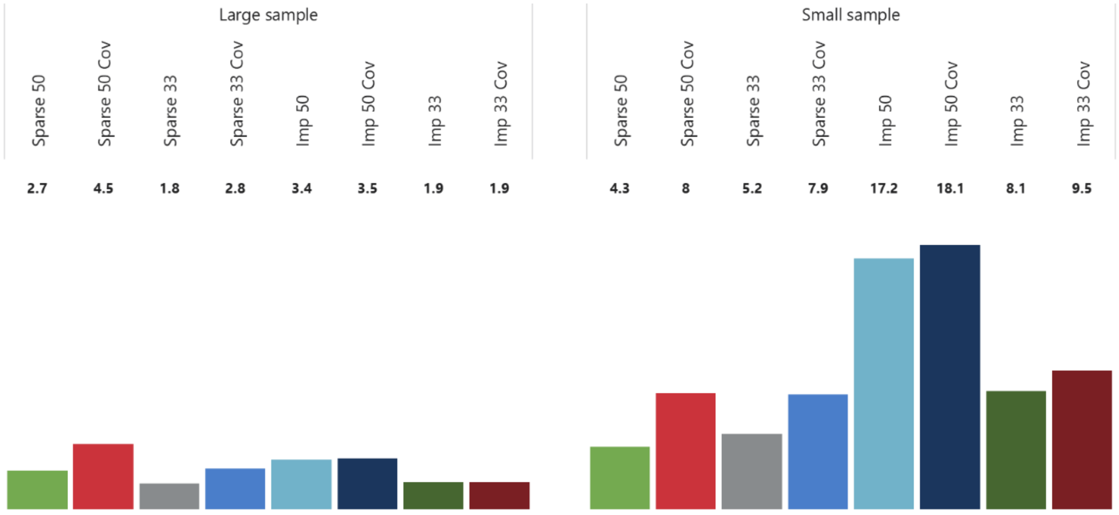
It is interesting to note that in the MAE analysis, the sparse 33 experimental cells do better than the benchmark, though it is unclear as to why this is the case.

Figure 10



MAE analysis was further conducted against the out-of-sample data sets (Figure 11). For the large data sets, it is again the sparse 33 experimental cells that perform the best, although the imputed 33 experimental cells also perform very favourably. However, in the smaller data sets the errors are much larger, indicating that during the HB estimation there is likely to have been a degree of overfitting, particularly with the imputed 50 experimental cells.

Figure 11



CONCLUSIONS/RECOMMENDATIONS

The primary goal of the research was to understand whether supplementing sparse designs with additional tasks, obtained through imputation, could provide sufficiently robust results, compared to a full design. It was hypothesized that imputed designs would outperform a sparse design as additional data would be included in the HB estimation. While some of the data would be inaccurate, due to HB's own imputation-like method for smoothing respondent data, it was believed that this would offset the inaccurate data and therefore would be superior.

This, however, proved not to be the case, and across all diagnostics and analysis the sparse designs outperformed the imputed designs. It should be noted however, that the original survey had not been set up with imputation specifically at mind. Many of the covariates in the original survey were categorical in nature, which meant that correlations with the conjoint tasks was low (typically, correlations of between 0.05-0.2 were observed). In addition, the covariates were not necessarily related to the attributes in the conjoint design so if it was known that imputation would be applied, more specific attribute-related questions could have been included to ensure a higher level of correlation.

The key takeaways from the research were:

- Even under favourable conditions for imputation, accuracy of predicting the correct concept struggled to exceed 50%.
- The number of respondents per version should be greater than 200 if imputation is to be applied to ensure stability of hit rates.
- Exceeding 300 respondents per version does not necessarily improve imputation success.
- The inclusion of covariates had minimal effect on goodness of fit, hit rates, and MAE, and caused significant overfitting issues on the data sets that contained N=500 respondents.
- Sparse designs outperformed imputed designs regarding in-sample validity.
- For out-of-sample predictions, the sparse and imputed designs performed equally well.
- Respondent-level correlations were extremely high even under extreme sparse conditions. For the imputed cells, the correlations were much lower.

Overall, the sparse 33 experimental cells were the clear winner when employed with the large data sets.

FURTHER RESEARCH ON IMPUTATION

The results of the imputation were mixed, which was primarily due to the categorical nature of the conjoint data. It is hypothesized that imputation will have better success when the choice is not categorical. Examples of this could be CVA style conjoint studies where respondents rate concepts on a numeric scale, or volumetric conjoint where respondents are asked how many of each concept they would purchase. It would also be of interest to understand the effects on hit rate where questionnaires have been set up with imputation in mind. Further investigation is also needed to understand if there are more appropriate imputation methods for categorical data.

During the Q&A session at the conference, it was commented that under no circumstances should an imputed design outperform a sparse design, but the authors do not necessarily believe that to be the case. What the research has shown is that adding in c.50% correct data and c.50% incorrect data leads to worse results, which shows that bad data hurts the analysis more than good data helps the analysis, but there must be a ratio of good to bad data which leads to better results for the imputed design. What that ratio is, and whether it is possible for imputation to achieve that level of accuracy is unknown and could be an area for further research.

Chris Moore

Ioannis Tsalamanis

APPENDIX

The diagnostics for each of the data sets can be found below.

Data Set 1

	Complete Large	Sparse Large 50	Sparse Large 50 Cov	Sparse Large 33	Sparse Large 33 Cov	Imp Large 50	Imp Large 50 Cov	Imp Large 33	Imp Large 33 Cov
	1	3	4	5	6	11	12	13	14
Per. Cert	713	745	799	723	753	605	618	636	647
RLH	630	665	724	640	673	530	541	557	567
Avg. Var	3.57	4.32	6.73	3.63	4.59	2.12	2.19	2.31	2.61
Agg correlation - All	-	0.98	0.99	0.99	0.99	0.98	0.98	0.99	0.99
Ind correlation - All	-	0.94	0.92	0.96	0.95	0.88	0.88	0.93	0.93
OOS MAE (unscaled)	3.8%	4.3%	4.1%	4.0%	4.1%	3.8%	3.7%	3.8%	3.8%
OOS MAE (scaled)	3.1%	3.0%	3.1%	3.0%	3.0%	3.2%	3.2%	3.1%	3.1%
HO - HR	66.9%	65.0%	62.9%	65.4%	64.9%	59.0%	58.8%	64.3%	64.3%
HO - MAE (unscaled)	4.3%	4.9%	4.6%	4.6%	4.6%	4.4%	4.5%	4.0%	4.0%
HO - MAE (scaled)	3.8%	3.6%	3.6%	3.7%	3.8%	3.9%	4.0%	3.6%	3.7%
Change in rank	-	3	6	2	2	5	5	2	2

	Complete Small	Sparse Small 50	Sparse Small 50 Cov	Sparse Small 33	Sparse Small 33 Cov	Imp Small 50	imp Small 50 Cov	imp Small 33	Imp Small 33 Cov
	2	7	8	9	10	15	16	17	18
Per. Cert	729	761	890	743	825	431	461	485	535
RLH	647	681	837	661	754	400	420	437	473
Avg. Var	3.67	4.20	27.90	4.08	9.00	0.74	0.99	0.91	1.324
Agg correlation - Small	-	0.95	0.84	0.97	0.96	0.96	0.96	0.95	0.96
Ind correlation - Small	-	0.91	0.81	0.95	0.90	0.69	0.67	0.78	0.76
OOS MAE (unscaled)	4.3%	5.0%	4.9%	4.7%	4.2%	6.6%	6.8%	5.8%	5.9%
OOS MAE (scaled)	2.9%	3.1%	3.3%	3.2%	3.0%	3.4%	3.4%	3.1%	3.2%
HO - HR	69.0%	64.6%	61.0%	67.0%	64.2%	50.6%	49.8%	54.6%	53.4%
HO - MAE (unscaled)	5.2%	5.3%	6.0%	5.3%	5.4%	9.1%	9.2%	8.6%	8.7%
HO - MAE (scaled)	4.7%	4.7%	5.2%	4.5%	5.1%	5.5%	5.5%	5.9%	5.9%
Change in rank	-	8	11	6	8	2	2	2	2

Data Set 2

	Complete Large	Sparse Large 50	Sparse Large 50 Cov	Sparse Large 33	Sparse Large 33 Cov	Imp Large 50	Imp Large 50 Cov	Imp Large 33	Imp Large 33 Cov
	1	3	4	5	6	11	12	13	14
Per. Cert	671	675	771	670	725	439	450	511	535
RLH	589	593	691	588	642	406	413	455	473
Avg. Var	2.54	2.77	6.01	2.60	3.80	0.93	0.96	1.20	1.37
Agg correlation - All	-	1.00	0.99	1.00	1.00	0.97	0.97	0.99	0.99
Ind correlation - All	-	0.84	0.79	0.92	0.90	0.67	0.65	0.83	0.81
OOS MAE (unscaled)	3.0%	4.3%	4.1%	3.7%	3.7%	2.9%	3.0%	3.3%	3.2%
OOS MAE (scaled)	2.6%	2.9%	3.1%	3.0%	3.0%	2.8%	2.9%	2.8%	2.8%
HO - HR	-	-	-	-	-	-	-	-	-
HO - MAE (unscaled)	-	-	-	-	-	-	-	-	-
HO - MAE (scaled)	-	-	-	-	-	-	-	-	-
Change in rank	-	0	4	2	2	6	6	2	2

	Complete Small	Sparse Small 50	Sparse Small 50 Cov	Sparse Small 33	Sparse Small 33 Cov	Imp Small 50	imp Small 50 Cov	imp Small 33	Imp Small 33 Cov
	2	7	8	9	10	15	16	17	18
Per. Cert	667	695	864	661	761	504	536	533	569
RLH	585	613	804	580	681	450	474	472	499
Avg. Var	2.65	3.07	19.76	2.50	6.09	1.46	1.89	1.46	2.04
Agg correlation - Small	-	0.98	0.98	0.99	0.99	0.96	0.96	0.98	0.97
Ind correlation - Small	-	0.81	0.72	0.91	0.85	0.61	0.58	0.79	0.76
OOS MAE (unscaled)	3.5%	5.5%	4.1%	4.2%	4.1%	4.1%	4.2%	3.7%	3.7%
OOS MAE (scaled)	2.9%	3.2%	3.3%	3.3%	3.5%	3.4%	3.4%	3.4%	3.4%
HO - HR	-	-	-	-	-	-	-	-	-
HO - MAE (unscaled)	-	-	-	-	-	-	-	-	-
HO - MAE (scaled)	-	-	-	-	-	-	-	-	-
Change in rank	-	7	2	4	8	7	7	8	6

Data Set 3

	Complete Large	Sparse Large 50	Sparse Large 50 Cov	Sparse Large 33	Sparse Large 33 Cov	Imp Large 50	Imp Large 50 Cov	Imp Large 33	Imp Large 33 Cov
	1	3	4	5	6	11	12	13	14
Per. Cert	757	770	850	767	820	556	590	636	655
RLH	676	691	785	687	748	490	517	557	574
Avg. Var	4.65	4.75	11.54	5.11	8.60	1.32	1.59	1.96	2.15
Agg correlation - All	-	0.99	0.99	1.00	0.99	0.98	0.98	0.99	0.99
Ind correlation - All	-	0.88	0.85	0.94	0.92	0.75	0.74	0.87	0.86
OOS MAE (unscaled)	4.3%	5.2%	4.9%	4.6%	4.6%	4.2%	4.1%	4.5%	4.5%
OOS MAE (scaled)	3.9%	4.2%	4.3%	3.9%	3.9%	3.9%	4.0%	4.0%	4.0%
HO - HR	65.4%	57.3%	55.6%	61.4%	61.3%	44.7%	43.4%	56.2%	55.3%
HO - MAE (unscaled)	2.2%	2.3%	2.9%	2.0%	1.9%	7.1%	7.1%	2.5%	2.3%
HO - MAE (scaled)	2.0%	2.3%	2.9%	1.5%	0.8%	5.8%	5.7%	2.5%	2.3%
Change in rank	-	7	9	2	5	8	10	9	9

	Complete Small	Sparse Small 50	Sparse Small 50 Cov	Sparse Small 33	Sparse Small 33 Cov	Imp Small 50	imp Small 50 Cov	imp Small 33	Imp Small 33 Cov
	2	7	8	9	10	15	16	17	18
Per. Cert	779	800	881	791	855	576	609	665	712
RLH	699	725	825	714	792	505	533	583	630
Avg. Var	5.59	5.79	19.88	5.82	12.32	1.69	2.19	2.51	3.64
Agg correlation - Small	-	0.98	0.96	0.99	0.99	0.96	0.96	0.99	0.99
Ind correlation - Small	-	0.85	0.77	0.93	0.88	0.67	0.65	0.85	0.82
OOS MAE (unscaled)	4.9%	5.6%	4.9%	5.2%	4.9%	4.9%	4.8%	4.9%	4.9%
OOS MAE (scaled)	4.3%	4.2%	4.4%	4.1%	4.4%	4.4%	4.4%	4.3%	4.4%
HO - HR	65.8%	54.4%	50.0%	63.0%	60.4%	39.0%	39.2%	52.6%	52.0%
HO - MAE (unscaled)	2.4%	4.8%	7.4%	3.5%	2.7%	5.8%	5.4%	3.1%	3.1%
HO - MAE (scaled)	2.3%	4.0%	4.0%	3.0%	2.4%	4.6%	4.5%	3.0%	3.1%
Change in rank	-	9	15	6	7	13	13	6	4

Data Set 4

	Complete Large	Sparse Large 50	Sparse Large 50 Cov	Sparse Large 33	Sparse Large 33 Cov	Imp Large 50	Imp Large 50 Cov	Imp Large 33	Imp Large 33 Cov
	1	3	4	5	6	11	12	13	14
Per. Cert	731	776	809	754	787	593	614	617	636
RLH	688	733	767	711	744	569	586	588	604
Avg. Var	3.55	4.64	6.58	4.16	5.98	1.95	2.31	2.06	2.40
Agg correlation - All	-	0.99	0.96	1.00	0.99	0.99	0.97	0.99	0.97
Ind correlation - All	-	0.88	0.76	0.94	0.88	0.74	0.71	0.85	0.81
OOS MAE (unscaled)	5.4%	5.4%	5.4%	5.4%	5.3%	5.5%	5.4%	5.6%	5.6%
OOS MAE (scaled)	5.4%	5.3%	5.3%	5.3%	5.3%	5.5%	5.4%	5.6%	5.6%
HO - HR	61.3%	55.4%	56.8%	59.5%	60.4%	49.6%	50.0%	55.2%	56.0%
HO - MAE (unscaled)	2.3%	2.7%	2.0%	2.4%	1.4%	3.2%	2.8%	2.2%	1.7%
HO - MAE (scaled)	1.5%	1.8%	1.6%	1.5%	1.1%	2.0%	1.6%	1.7%	1.2%
Change in rank	-	2	6	4	5	2	2	2	2

	Complete Small	Sparse Small 50	Sparse Small 50 Cov	Sparse Small 33	Sparse Small 33 Cov	Imp Small 50	imp Small 50 Cov	imp Small 33	Imp Small 33 Cov
	2	7	8	9	10	15	16	17	18
Per. Cert	726	781	863	753	835	601	622	647	689
RLH	684	738	827	710	796	575	592	613	649
Avg. Var	3.53	2.50	13.50	4.48	9.81	2.28	2.92	2.72	3.79
Agg correlation - Small	-	0.97	0.92	0.98	0.96	0.89	0.87	0.97	0.97
Ind correlation - Small	-	0.85	0.70	0.91	0.84	0.60	0.56	0.82	0.78
OOS MAE (unscaled)	5.7%	5.9%	5.7%	5.8%	5.7%	6.4%	6.4%	6.1%	6.1%
OOS MAE (scaled)	5.6%	5.4%	5.5%	5.5%	5.6%	6.4%	6.4%	6.0%	6.0%
HO - HR	59.8%	56.2%	54.0%	57.2%	57.2%	48.8%	47.2%	53.2%	55.8%
HO - MAE (unscaled)	2.9%	3.6%	4.4%	3.6%	3.0%	2.6%	2.3%	3.7%	3.4%
HO - MAE (scaled)	2.7%	2.7%	3.3%	2.9%	2.5%	2.2%	2.1%	2.2%	2.2%
Change in rank	-	8	11	6	7	10	12	9	9

Data Set 5

	Complete Large	Sparse Large 50	Sparse Large 50 Cov	Sparse Large 33	Sparse Large 33 Cov	Imp Large 50	Imp Large 50 Cov	Imp Large 33	Imp Large 33 Cov
	1	3	4	5	6	11	12	13	14
Per. Cert	707	732	758	720	722	613	613	650	636
RLH	666	690	715	678	680	585	585	616	604
Avg. Var	3.98	4.64	5.95	4.28	4.37	2.48	2.28	2.85	2.61
Agg correlation - All	-	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99
Ind correlation - All	-	0.86	0.83	0.93	0.92	0.72	0.70	0.87	0.86
OOS MAE (unscaled)	5.4%	6.0%	5.6%	5.4%	5.5%	5.1%	5.1%	5.3%	5.2%
OOS MAE (scaled)	3.9%	3.8%	3.7%	3.8%	3.9%	3.9%	3.9%	3.6%	3.7%
HO - HR	59.2%	51.1%	48.8%	57.0%	57.1%	46.8%	47.3%	55.3%	55.7%
HO - MAE (unscaled)	4.3%	6.0%	8.5%	4.8%	4.9%	3.6%	3.4%	2.9%	2.7%
HO - MAE (scaled)	1.8%	2.0%	2.5%	1.8%	1.6%	2.9%	2.6%	2.4%	2.1%
Change in rank	-	2	4	5	2	2	2	5	5

	Complete Small	Sparse Small 50	Sparse Small 50 Cov	Sparse Small 33	Sparse Small 33 Cov	Imp Small 50	imp Small 50 Cov	imp Small 33	Imp Small 33 Cov
	2	7	8	9	10	15	16	17	18
Per. Cert	721	761	806	717	767	678	716	666	687
RLH	679	718	764	675	724	640	675	629	648
Avg. Var	3.99	7.16	10.07	2.12	6.90	3.87	5.60	3.06	3.92
Agg correlation - Small	-	0.96	0.96	0.99	0.99	0.86	0.86	0.96	0.97
Ind correlation - Small	-	0.76	0.71	0.92	0.88	0.59	0.57	0.83	0.81
OOS MAE (unscaled)	5.5%	6.3%	6.0%	6.1%	6.3%	7.2%	7.4%	6.0%	6.0%
OOS MAE (scaled)	4.0%	4.5%	4.5%	4.3%	4.5%	5.4%	5.5%	4.4%	4.4%
HO - HR	58.4%	50.0%	53.4%	57.6%	58.6%	46.2%	47.6%	53.6%	53.6%
HO - MAE (unscaled)	2.9%	5.1%	3.8%	3.8%	2.8%	3.9%	4.2%	2.5%	2.4%
HO - MAE (scaled)	2.2%	1.9%	1.6%	2.2%	1.7%	3.0%	2.5%	1.7%	1.5%
Change in rank	-	8	7	2	4	9	8	4	5

References:

Moore, Chris et al. (2016), "Choice Based Conjoint in a Mobile World—How far can we go," Sawtooth Software proceedings 2016, pp 97–116.

Orme, Bryan & Howell, John (2009), "Application of Covariates Within Sawtooth Software's CBC/HB Program: Theory and Practical Example," Sawtooth Software Website.

Sentis, Keith & Geller, Valerie (2010), "The impact of Covariates on HB Estimates." Sawtooth Software proceedings 2010, pp 255–268.

Johnson, Richard & Orme, Bryan (1996), "How Many Questions Should You Ask in Choice-Based-Conjoint Studies," Sawtooth Software Website.

COMBINING CHOICE-BASED CONJOINT AND DYNAMIC CHOICE MODELS FOR MORE ACCURATE FORECASTING

FAINA SHMULYIAN
SKIM USA

ABSTRACT

Sales forecasting for expensive and innovative products using Choice-Based Conjoint (CBC) requires extra attention to changes in consumer behavior over time. This paper demonstrates how to incorporate the ideas of the Bass Diffusion Model (BDM) and Evidence-Accumulation models with a CBC and presents a new way of introducing a dynamic element directly into individual choice models. Major advantages of a traditional CBC and different classes of dynamic models are combined in the new approach.

INTRODUCTION

Forecasting using the Choice-Based Conjoint (CBC) model is one of the most common approaches for evaluating the potential of a new product. Researchers favor CBC because it can be used when little or no sales history is available. It also allows accurate simulations of hypothetical scenarios. Finally, the CBC can take into account market heterogeneity.

One of the limitations of forecasting with a CBC is that consumer utilities are assumed to be constant over time. This assumption may not be problematic for a simple Consumer Packaged Goods (CPG) category, such as dish detergent, where patterns of consumer behavior are stable, decisions are made almost instantaneously, and costs associated with a wrong decision are relatively low.

The assumption in the CBC that consumer utilities are static, however, can introduce significant forecasting error when researchers are analyzing complex and expensive categories, such as technology and luxury goods. Consumer evaluation of these innovative products is greatly influenced by temporal factors. Forecasting using only the CBC will be less accurate when consumer decisions take more time (such as with technologically complex products) and consumers face higher risks for wrong choices (such as with high cost products).

INTERNAL AND EXTERNAL TEMPORAL FACTORS

There are multiple internal factors within consumer cognition that can change over time. One of the most important is what consumers do and do not know about a product's attributes. Product knowledge may be quite simple for dish detergent but much more complex for a cell phone. Even a moderately simple product, such as smart-home technology, may challenge consumers' understanding if they are considering a new product that has just been introduced to the market. In this purchasing context, consumers might choose a product based on only a subset of the attributes rather than all attributes of the product. This subset of attributes can change over time as consumers gain greater knowledge of product complexity. Consumer knowledge is especially likely to change if buyers start with little or no prior knowledge or well-defined preferences of a product, which is quite common for complex and innovative products.

There are also multiple external factors that alter consumer preferences over time. Changes in the economic circumstances of a household or lifestyle changes (e.g., buying a home or the birth of a child) will almost certainly reshape purchasing behavior. Changes in popular culture, fashion, and advertising are constantly altering what consumers want. In addition, consumers often imitate each other so that purchases by family members, peers, and friends have a feedback effect on the purchasing preferences of other people in the social network. This effect is often termed "word-of-mouth."

Temporal Forecasting Using the Bass Diffusion Model

The Bass Diffusion Model (BDM) is one of the most popular dynamic models for forecasting [1,2]. It conceptualizes a "product diffusion process" to measure how word-of-mouth influences sales over time. BDM is particularly valuable for forecasting long-term sales patterns for new technologies and durable goods. It has been shown to reliably predict the timing and volume of new product adoptions (first purchases).

BDM makes two major assumptions. First, many products have a generic pattern of temporal penetration into the consumer base that occurs over months or even years. Second, the main drivers of penetration for a new product are innovation (such as advertising) and imitation (word-of-mouth).

Given these assumptions, the basic BDM equation is:

$$L(t) = p+qS(t),$$

where

$L(t)$ is the conditional likelihood that a customer will adopt the innovation at exactly time since introduction,

t is time,

$S(t)$ is the share of consumers who have already adopted the innovation by time t,

p is the coefficient of innovation, and

q is the coefficient of imitation.

The coefficient of innovation and imitation are usually estimated based on historic sales data or surveys of consumer purchase intentions.

While BDM does add temporal variables to product forecasting, it still faces three problems related to parametrization. First, is how to make predictions for an entirely new product. While the use of analogs can help firms make forecasts before introducing innovation into a market, suitable analogs might not be available. Second, most of the historic data from analogs describe how successful innovations diffuse through the population, which can introduce success bias into forecast if the new product does not succeed in the market. Third, accurate parameter estimation is only possible after making several observations of actual sales for the new products, but by this time a firm has often already made critical investment decisions.

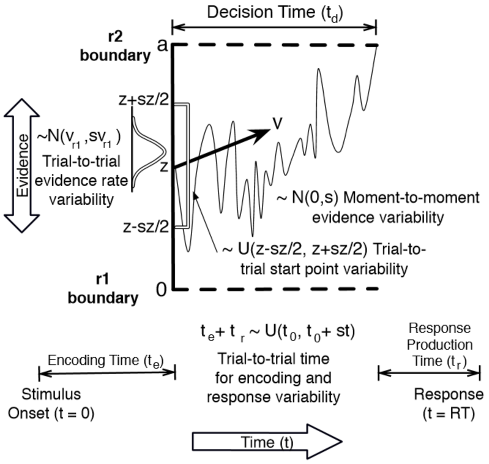
In addition to parametrization, BDM is further limited by multiple internal and external factors. First, classic BDMs, and even modern extensions, don't consider multiple product attributes and usually make predictions on a sub-category level. Second, most BDMs don't account for changing market scenarios. Third, BDMs don't measure population heterogeneity since they are designed to estimate average demand. Fourth, the basic BDM does not incorporate effects of advertising and other external factors that can change the parameters over time.

Despite these limitations, BDM can successfully complement and enhance a CBC for more accurate forecasting. To realize this potential synergy requires incorporating the diffusion process directly into individual choice models for each respondent.

COMBINING INDIVIDUAL CHOICE AND THE DIFFUSION PROCESS THROUGH EVIDENCE-ACCUMULATION MODELS

Psychology and sociology use the concept "evidence accumulation" to describe how individuals move from little or no knowledge of a topic to partial or even complete knowledge. The Diffusion Decision Model (hereafter DDM) is a popular way to measure this process for consumers [8]. The DDM assumes that each individual is making a binary choice to buy or not to buy a product. It models an individual's positive or negative choice as a continuous, stochastic process. The probability of a consumer purchasing a product at a given point in time is based on the accumulation of positive and negative evidence. Only when the accumulation of positive evidence reaches a certain threshold for a consumer will she make a positive purchasing decision (see Fig.1).

Figure 1: Diffusion Decision Model


Decision Time (t_d)
r2 boundary
a
z+sz/2
Evidence
~N(v_{r1}, sv_{r1})
Trial-to-trial evidence rate variability
z
v
~ N(0,s) Moment-to-moment evidence variability
z-sz/2
~ U(z-sz/2, z+sz/2) Trial-to-trial start point variability
r1 boundary
0
$t_e + t_r \sim U(t_0, t_0 + st)$
Trial-to-trial time for encoding and response variability
Encoding Time (t_e)
Stimulus Onset (t = 0)
Response Production Time (t_r)
Response (t = RT)
Time (t)

Another threshold-based evidence-accumulation model was suggested by Granovetter (1978). According to Granovetter, consumer purchases are a form of collective behavior in which the utilities of alternatives (benefits and costs) depend on how many other actors choose that alternative. Some proportion of other consumers must make a positive decision before a given actor will also do so. This threshold point is where the net benefits begin to exceed the net costs to an actor, such as the introduction of cell phones which eventually reaches a tipping point so that consumers lacking a cell phone feel "left out." If a researcher knows the distribution of individual thresholds, she can estimate the equilibrium share of actors making each decision.

The suggested dynamic utility model is incorporating principles of the BDM and evidence-accumulation models summarized above. The assumption is that when a product is just introduced to the market, consumers are relying only on their impression of the product when they consider buying it. This impression is assessed by building a CBC model and estimating individual utilities in a current market scenario with the new product included into the consideration set. Even if a consumer likes the new product (has a high utility for it), she might not be the one to be driven by innovation and will not buy it as soon as possible. Other consumers might prefer purchasing something new and "cool" instead of a familiar product on the market. In the suggested model the difference is expressed by introducing an innovation coefficient similarly to the BDM but for each respondent individually. After some consumers purchase the new product, those who are still considering buying it will start accumulating evidence that would affect their interest in the product. They will read reviews on social media, hear from their family and friends, and see people using this product in different situations. In the model, the evidence-accumulation will be represented by additional term proportional to the preference share estimated in the CBC at the previous time step. The higher the new product share, the more potential consumers are exposed to the product, the more evidence is accumulated. Again, the additional term will be weighted like in the BDM. Some consumers are more affected by the word-of-mouth and other interactions with the product than others. The individual imitation coefficient expresses this difference. As a result, similarly to the Granovetter model, individual part-worths of the new product will consist of part-worths of its attributes and the dynamic term proportional to the number of consumers who have already purchased this product. In the CBC context, the threshold will be represented by the part-worth of alternatives and the None option in the simulated scenario.

The Dynamic Utility Model

The suggested dynamic model of product utility (that is, the combination of CBC, BDM, and DDM) is based on the following interpretation of a product part-worths:

$$W_i^k(t) = p_i V_i^k(t) + q_i U_i^k(t),$$

where

$W_i^k(t)$ is the part-worth of a product *k* for a respondent *i* at time *t*,

p_i is the coefficient of innovation,

q_i is the coefficient of imitation,

$V_i^k(t)$ is the product part-worth as usually estimated in a CBC, and

$U_i^k(t)$ is the utility of a market share of a product *k* as perceived by a person *i* at time *t*.

The utility of market share can be expressed as follows:

$$U_i^k(t) = c_i^k * r_i^k * SP^k(t-1),$$

where

$U_i^k(t)$ is the utility of a market share of a product *k* as perceived by a person *i* at time *t,*

c_i^k is the contact rate of individual *i* with people who adopted alternative *k,*

r_i^k is the recommendation rate of alternative *k* received by individual *i,* and

$SP^k(t-1)$ is the share of a product *k* at the previous moment of time.

There are several important assumptions built into this dynamic utility model. First, product attribute utilities don't change over time. It is also assumed that innovation and imitation coefficients (p_i and q_i*)* differ for different respondents, don't change over time, and can be estimated based on survey data. Contact and recommendation rates (*c* and *r)* are calculated using stated brand contact and recommendation numbers for each respondent. It is assumed that both *c* and *r* are constant over time. These assumptions won't significantly affect the predictive powers of the model if these internal factors don't change as fast as external factors, such as growth in interest due to the word-of-mouth. Another assumption is that the effects of a negative word-of-mouth are not simulated. Finally, advertisement effects are not considered in a basic model but could be included using an extension of BDM:

$$VS_i^k(t) = \beta AI^k(t)[pV_i^k(t) + qU_i^k(t)],$$

where

$AI^k(t)$ is the increase in advertisement in a period of time prior to t, and

β is the coefficient capturing the percentage increase in diffusion speed resulting from a 1% increase in advertising.

Implementing the dynamic utility model for a product requires completing the following five steps.

1. Conduct a standard choice exercise and estimate products part-worths to utilize the advantages of CBC forecasting.
2. Use a "Triple-response None" follow-up in the CBC exercise to estimate each respondent's potential to adopt/buy alternatives in the future (utility of "Maybe"). The choice question is asked the same way as for a Dual None except that respondents are asked to select one of three options: "I will buy this product," "I am considering buying this product, but I have not decided yet," and "I will not buy this product." The individual utilities of "Maybe" as an alternative-specific attribute can be estimated simultaneously with all other utilities in the CBC model. The dynamic model redistributes "Maybe" between "Yes" and "No" over time for each respondent.
3. Incorporate the effect of the word-of-mouth directly into each individual choice model by adding a utility for a perceived market share to a product utility following the structure of a BDM. The initial share of preference of a product or portfolio in a given scenario is estimated using a traditional CBC model. Additional survey questions about the number of contacts and recommendations are used to calculate the contact and recommendation rates for each respondent.
4. Estimate the innovation and imitation coefficients individually for each respondent. The CBC utilities and respondents' answers to questions assessing how others are influencing their decisions are used as explanatory variables. Bayesian regression is applied to estimate the impact of CBC attributes and external parameters on respondents' decisions. Hierarchical Bayes regularizes the estimation since all parameters are drawn from the same population distribution. The hierarchical model uses the information from an entire group captured in the population-level parameters to improve parameter estimation at the individual level.
5. As an option, calibrate the model with historic data if it is available and relevant.

The individual parameters required for the model and estimated outside of the CBC are summarized in Table 1.

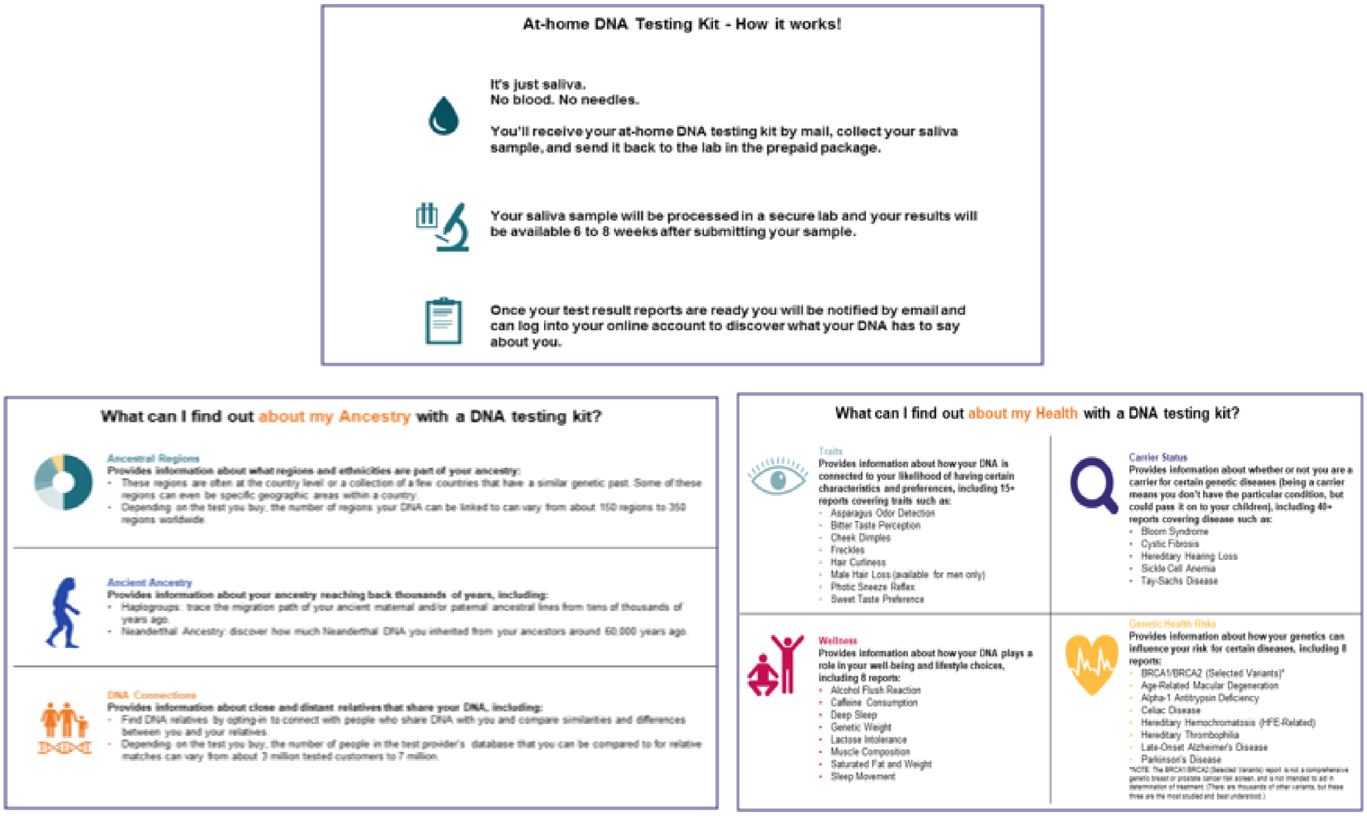
Table 1: Parameters of the Dynamic Model Estimated Outside of the CBC

p_i—coefficient of innovation	An agreement rating was collected to a battery of behavioral questions related to the new technology ("I try to be the first to buy new gadgets," "I make sure I read reviews before I buy a new product," "I often ask my family and friends for opinion about a new product," etc.). Based on these ratings, p_i was calculated as a propensity score of being an "innovator."
q_i—coefficient of imitation	In the suggested model, we assume that $q_i = 1 - p_i$. In general, the two coefficients can be independent and estimated separately.
c_i^k—contact rate	The stated number of contacts for the brand for each respondent in the past six months was used to estimate the contact rate.
r_i^k—recommendation rate	The stated number of recommendations for the brand in the past six months was used to estimate the recommendation rate.

Case Study Simulation

The author worked with a leading company in genetic testing to design and implement a marketing research study. The innovative product to assess was an at-home DNA testing kit. The company was expanding its portfolio considering multiple bundles of report categories in Ancestry and Health (see Figure 2).

Figure 2: Educational pages from the survey explaining the company's report categories to respondents.



At-home DNA Testing Kit - How it works!

It's just saliva.
No blood. No needles.

You'll receive your at-home DNA testing kit by mail, collect your saliva sample, and send it back to the lab in the prepaid package.

Your saliva sample will be processed in a secure lab and your results will be available 6 to 8 weeks after submitting your sample.

Once your test result reports are ready you will be notified by email and can log into your online account to discover what your DNA has to say about you.

What can I find out about my Ancestry with a DNA testing kit?

Ancestral Regions
Provides information about what regions and ethnicities are part of your ancestry:
- These regions are often at the country level or a collection of a few countries that have a similar genetic past. Some of these regions can even be specific geographic areas within a country.
- Depending on the test you buy, the number of regions your DNA can be linked to can vary from about 150 regions to 350 regions worldwide.

Ancient Ancestry
Provides information about your ancestry reaching back thousands of years, including:
- Haplogroups: trace the migration path of your ancient maternal and/or paternal ancestral lines from tens of thousands of years ago.
- Neanderthal Ancestry: discover how much Neanderthal DNA you inherited from your ancestors around 60,000 years ago.

DNA Connections
Provides information about close and distant relatives that share your DNA, including:
- Find DNA relatives by opting-in to connect with people who share DNA with you and compare similarities and differences between you and your relatives.
- Depending on the test you buy, the number of people in the test provider's database that you can be compared to for relative matches can vary from about 3 million tested customers to 7 million.

What can I find out about my Health with a DNA testing kit?

Traits
Provides information about how your DNA is connected to your likelihood of having certain characteristics and preferences, including 15+ reports covering traits such as:
- Asparagus Odor Detection
- Bitter Taste Perception
- Cheek Dimples
- Freckles
- Hair Curliness
- Male Hair Loss (available for men only)
- Photic Sneeze Reflex
- Sweet Taste Preference

Carrier Status
Provides information about whether or not you are a carrier for certain genetic diseases (being a carrier means you don't have the particular condition, but could pass it on to your children), including 40+ reports covering disease such as:
- Bloom Syndrome
- Cystic Fibrosis
- Hereditary Hearing Loss
- Sickle Cell Anemia
- Tay-Sachs Disease

Wellness
Provides information about how your DNA plays a role in your well-being and lifestyle choices, including 8 reports:
- Alcohol Flush Reaction
- Caffeine Consumption
- Deep Sleep
- Genetic Weight
- Lactose Intolerance
- Muscle Composition
- Saturated Fat and Weight
- Sleep Movement

Genetic Health Risks
Provides information about how your genetics can influence your risk for certain diseases, including 8 reports:
- BRCA1/BRCA2 (Selected Variants)*
- Age-Related Macular Degeneration
- Alpha-1 Antitrypsin Deficiency
- Celiac Disease
- Hereditary Hemochromatosis (HFE-Related)
- Hereditary Thrombophilia
- Late-Onset Alzheimer's Disease
- Parkinson's Disease

*NOTE: The BRCA1/BRCA2 (Selected Variants) report is not a comprehensive genetic breast or prostate cancer risk screen, and is not intended to aid in determination of treatment. (There are thousands of other variants, but these three are the most studied and best understood.)

The study sought to answer three questions about portfolio and price optimization:

1. Which are the most valued packages across Ancestry and Health?
2. What would be the impact of a middle-tier introduction?
3. What would be the optimal price structure of this introduction in order to maximize customer reach and revenue?

The company emphasized that there were two challenges to optimizing their portfolio of DNA Test offerings. First, the genetic testing market was changing rapidly and its share was growing significantly every month. Second, word-of-mouth was a very important driver of a purchase in this category and must be considered in the modeling.

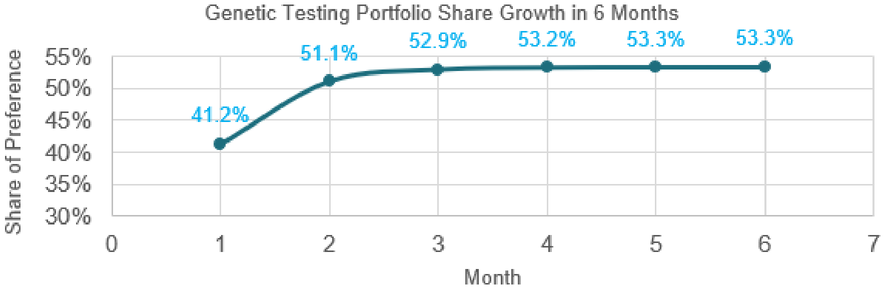
Respondents (N=2303) completed a survey where they answered a series of diagnostic questions and responded to a traditional CBC exercise. Respondents completed a series of screens where they made trade-offs between three different product offerings. On every screen, the attributes of each product offering varied, including ancestry features, health features, and price. The "Triple-response None" provided a benchmark on the overall appeal of the respondent's choice.

The modelling process followed the five steps summarized above. All the utilities were estimated using CBC HB, Sawtooth Lighthouse 9.3. The time step (Δt) of one month was selected for all simulations based on the current market dynamics in the category. To incorporate the dynamic effect of word-of-mouth into each individual model, we used

additional survey questions to measure the share of preference for a given product scenario and the contact and recommendation rates in the past six months reported by each respondent. The innovation and imitation coefficients were fitted using the HB procedure implemented in R ggdmc package:
https://www.rdocumentation.org/packages/ggdmc/versions/ 0.2.5.2

The resulting dynamic model predicted a 29% growth for the optimized portfolio in six months following the survey (see Figure 3) which is very close to the actual DNA database growth reported by the company during the same time period [9].

Figure 3. Six-Months Share Growth Prediction for the Optimized Genetic Testing Portfolio



Conclusion

Sales forecasting for durable, expensive, and innovative products is quite different from forecasting for products that are quickly used up, cheap, and low-tech. The primary difference is that complex products require measuring changes in consumer utility over time.

A case study of a DNA testing kits reveals the challenges of designing a marketing research project for such nontraditional products. The DNA testing market is relatively new and growing very rapidly. In addition, the kit has both ancestry and health attributes.

To address these challenges, the marketing research project was designed around the concept of "evidence accumulation." The study assumed that elements of both the CBC and Dynamic Choice Models needed to be included in order to estimate consumer preferences of a new middle-tier DNA kit.

The combined model successfully estimated the optimal price structure of this introduction and how the price would maximize customer reach and revenue. This exercise suggests six advantages of the combined model. First, it enables researchers to adjust models for changing market situations. Second, it can simulate "what if" scenarios. Third, the combined model can forecast for products with no or minimal sales data. Fifth, it can consider heterogeneity of consumer preferences. Sixth, the combined model can incorporate external factors, such as the word-of-mouth and advertising.

Faina Shmulyian

References

[1] F.M. Bass, "A New Product Growth Model for Consumer Durables," Management Science, Vol. 15, No.5, pp. 215–227, 1969.

[2] V. Mahajan, E. Muller and F.M. Bass, "New Product Diffusion Models in Marketing: A Review and Directions for Research," Journal of Marketing, Vol. 54, Issue 1, pp. 1–26, 1990.

[3] M. Granovetter, "Threshold Models of Collective Behavior," The American Journal of Sociology, Vol. 83, No.6, pp. 1420–1443, 1978.

[4] D. Horsky, "A Diffusion Model Incorporating Product Benefits, Price, Income and Information," Marketing Science, Vol. 9, No.4, pp 342–365, 1990.

[5] W.-J. Kim, J.-D. Lee and T.-Y. Kim, "Demand Forecasting for Multigenerational Products Combining Discrete Choice and Dynamic Diffusion Under Technological Trajectories," Technological Forecasting and Social Change, Vol. 72, pp. 825–849, 2005.

[6] J. Lee, Y. Cho, J.-D. Lee and C.-Y. Lee, "Forecasting Future Demand for Large Screen television Sets Using Conjoint Analysis and Diffusion Model," Technological Forecasting and Social Change, Vol. 73, pp. 362–376, 2006.

[7] S. Meeran, S. Jahanbin, P. Goodwin and J.Q.F. Neto, "When Do Changes in Consumer Preferences Make Forecasts From Choice-Based Conjoint Models Unreliable?," In Press, 2018.

[8] A. Heathcote, Y.-S. Lin, A. Reynolds, L. Stickland, M. Gretton, and D. Matzke, "Dynamic Models of Choice," In Press, Behavior Research Methods, 2018.

[9] L. Larkin, "DNA companies' databases growth in years 2012–2019," www.theDNAgeek.com/dna-tests, 2019

DATA FUSION: A FLEXIBLE HB TEMPLATE FOR MODELING STRUCTURES ACROSS MULTIPLE DATA SETS

KEVIN LATTERY
SKIM GROUP

1.0 INTRODUCTION

Many research studies use only one source of data, but in our age of expanding data we are more likely to find ourselves with two or more sources of data. In addition, we may need to make sense of these multiple data sources in relation to each other. This is an example of what we call data fusion. One way we can make sense of multiple data sets is qualitatively, by interpreting the results in relationship to each other and using expert knowledge to tell a cohesive story. This is a perfectly valid practice that researchers have been doing for decades. But we will not be discussing qualitative data fusion here. We will be discussing *quantitativ*e data fusion, where we make sense of multiple sets of data in relation to each other by quantitative analysis.

There are many types of data that we might want to analyze in relation to each other. The table below shows just a few examples.

Data Set 1	Data Set 2
Conjoint	Buy or Not Buy
MaxDiff	Anchor Question
Conjoint or MaxDiff	Purchase Intent Ratings
Conjoint	Ratings of Levels
Conjoint	MaxDiff/Q-Sort
Conjoint	Sales Data
Customer Data	Transaction Data
Marketing Spend	Regional Performance

We have deliberately included a few examples that might be familiar for users of Sawtooth Software. Their familiarity might lead one to overlook that they really are cases of data fusion. For example, a MaxDiff study with an anchor is a simple example of data fusion because we have two sources of data with two different types of questions.

The topic of data fusion is a broad one, and this paper makes no attempt to be exhaustive. Instead, we describe three general approaches to data fusion. We follow this with a technical treatment of the most complex approach, and apply that to some common cases.

The detailed case studies here all involve data fusion where the different data sources are directly linked because they have the same respondents. This is the kind of data fusion most often practiced by users of Sawtooth Software. Other kinds of data fusion, for instance between survey and aggregate behavioral data, will only briefly be touched upon when we discuss general approaches.

2.0 Three General Approaches to Data Fusion

One can approach any specific data fusion project with 3 general strategies. Given a specific problem one can choose an approach. In some cases, one or more of the approaches may be very difficult and a specific approach may be chosen for practical purposes. We call the three general approaches:

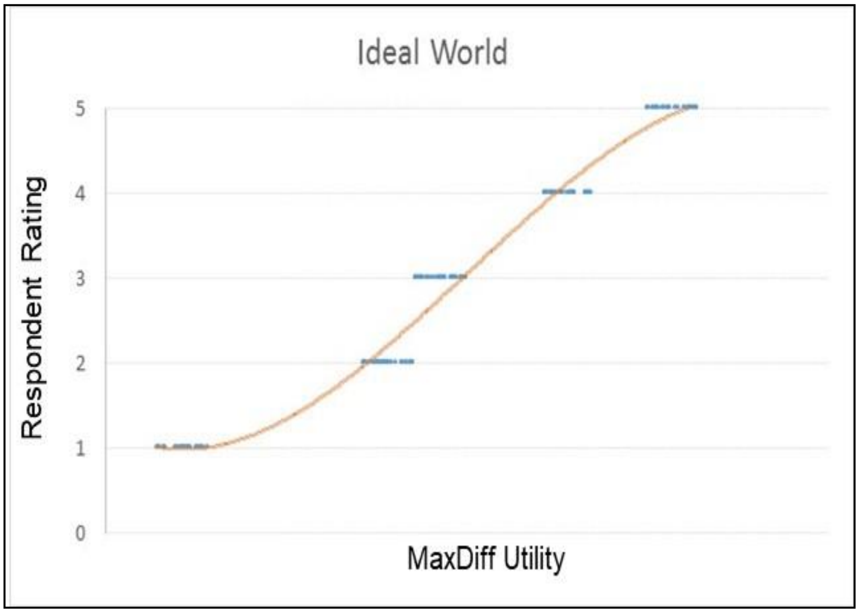
1. Two-Stage Linkage
2. Data Augmentation/Stacking
3. Complete Structural Model/Probabilistic Programming

We now describe each of these approaches separately.

2.1 Two-Stage Linkage

Two-stage Linkage is the easiest to understand. The idea is that we have a first stage of modeling that is just standard modeling, with no data fusion. We then have a second stage where we build a model that links the two data sources.

For example, if we have a conjoint and a MaxDiff study, the first stage analyzes each of them separately. We get conjoint results and MaxDiff results. The second stage is the difficult one, where we try to link the utilities together. Another example is we have MaxDiff data and purchase intent ratings. We analyze the conjoint data in the first stage. In the second stage, we try to link the MaxDiff utilities with the purchase intent rating scales. In an ideal world you might see something like this:



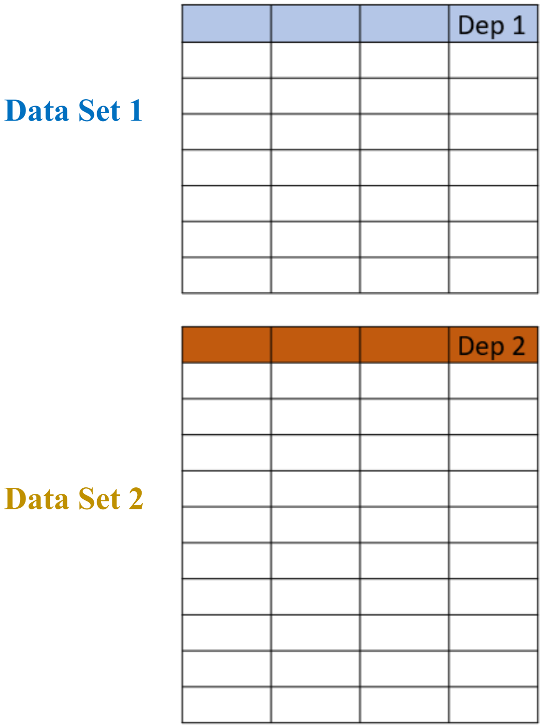
One can then fit a simple curve for the second stage linkage. Unfortunately this ideal world rarely happens with real data. We strongly caution the reader to avoid relying on finding such clear relationships like the chart above. More often we expect that one will need to apply some very clever modeling, and even then one may not be able to link the data well.

A better example of a two-stage approach is when we use the first stage as priors and then model a second set of data. For example, we might have conjoint data and related sales data over time. Our primary goal might be sales forecasting. In this context, one approach we have found worthwhile treats the analysis of conjoint data as the first stage. This is done without any reference to the sales data. We then use the conjoint results as informative priors for analyzing sales data over time. For instance, we might use the conjoint pricing elasticities and product switching as priors when we estimate the model for the sales data. Here the second stage estimation is making use of the first stage as priors that can be adjusted. This allows us to model the sales data over time, with the conjoint as a kind of constraint on the modeling of sales data that prevents overfitting.

The above example hints at something like a joint estimation because it analyzes the data in stage two jointly with stage one. But we still consider it as a two-stage method since the first stage is a standard independent analysis that ignores the other set of data. In contrast, the remaining two approaches analyze the two sets of data ***jointly***. When possible, we generally prefer a joint estimation over a two-stage sequential approach. That said, it is possible in a specific case that a two-stage approach may work better for specific goals than a joint estimation.

2.2 Data Augmentation/Stacking

Data Augmentation allows one to build a single model using both sets of data simultaneously. For users of Sawtooth Software, this is a common power trick. This approach to data fusion takes the two sets of data and makes them one data set by stacking them together.



For example, with a MaxDiff and conjoint, one would create a data file that has all the parameters from the union of the MaxDiff and conjoint. This means you will likely add parameters to both data sets. For example, parameters in data set 1 that are not in data set 2 must be added to data set 2. These added parameters are assigned a value of 0 as there is no information on them for that specific data set. So, the method requires some clever recoding, and in many cases there may be multiple ways to recode the data.

Over the years there have been many presentations at the Sawtooth Software Conference describing various methods of stacking multiple sets of data. At the end of this paper in the References section we list 8 papers on the data augmentation/stacking approach applied to various kinds of data fusion.

One advantage of data augmentation over a two-stage approach is that data augmentation is a joint estimation that considers all the data. In contrast, the two-stage approach does not consider all the data jointly, at least in the first stage. Another advantage of the data augmentation approach is that one does not have to do a second-stage linkage. Adding parameters and stacking, coupled with joint estimation, is the linkage. This can save time and be less frustrating than trying to link models on a second stage, especially if results from the first stage are not well aligned.

A practical disadvantage of the data augmentation approach is that the data file is larger. We now have additional tasks per respondent and more parameters (many of which might have values of 0). This makes estimation take longer. In addition, one should expect to need more MCMC iterations when doing full Bayesian estimation. The two data sets have different underlying structures that have been stacked together. This means MCMC iterations take longer to converge to a stationary distribution. We highly recommend testing convergence by running multiple chains and using a rigorous convergence test like Gelman-Rubin.

The key conceptual disadvantage of data augmentation is that the two sets of data most likely have different scales. Adjusting for this difference in scale between multiple data sources is a key component in doing data fusion well. The topic is so important it deserves its own section in this paper, which we turn to next.

2.2.1 Different Stimuli, Different Cognitive Processes, Different Scales

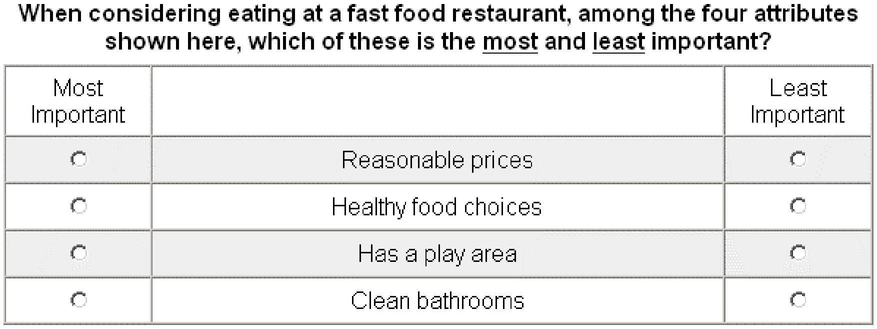
As researchers we observe responses to stimuli. The stimulus might be a specific survey question. For example, it might be a common rating scale question like this:

Please **rate** how important each of the following features is for a fast food restaurant:

[**Show n Scale Pts**]

1. Reasonable Prices
2. Healthy Food Choices
3. Has a Play Area
4. Clean Bathrooms

For a data fusion project we would have two sets of data with two (likely different) stimuli. For instance, we might also have a MaxDiff question like this:



When considering eating at a fast food restaurant, among the four attributes shown here, which of these is the most and least important?

Most Important		Least Important
○	Reasonable prices	○
○	Healthy food choices	○
○	Has a play area	○
○	Clean bathrooms	○

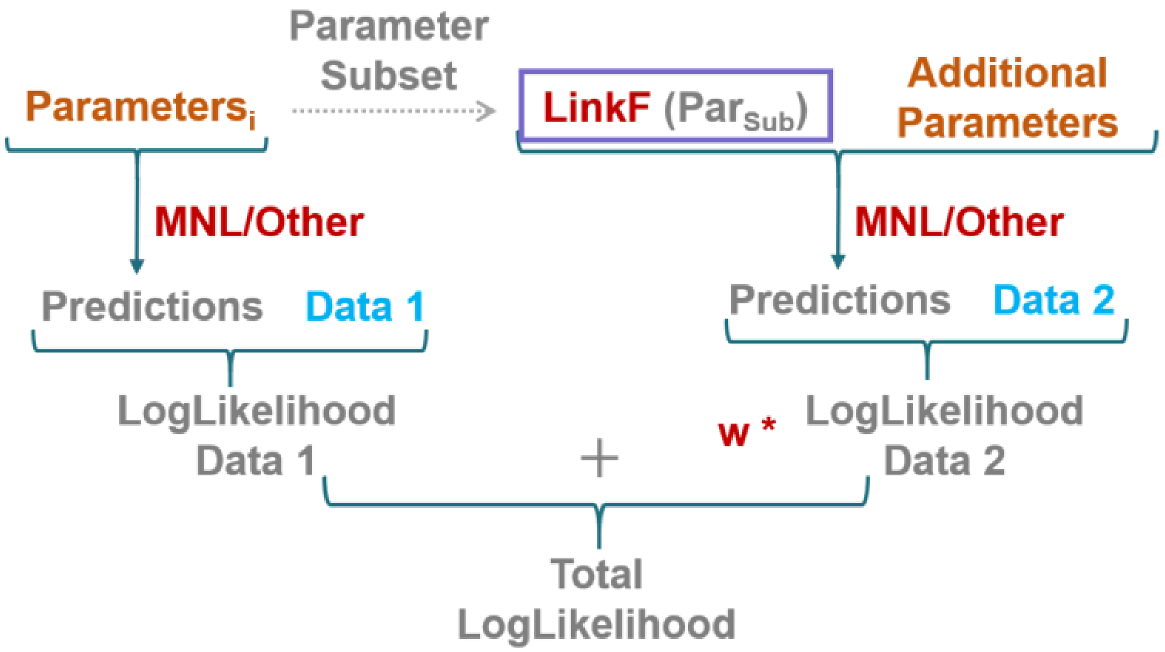
These two stimuli ask about the same four items. Regarding these four items, we think the respondent has some kind of underlying preferences. We don’t observe these underlying preferences. Instead we ask questions in the form of stimuli like the two different survey questions above. The respondent then uses cognitive processes to convert their true preferences to responses. In the first case we are asked to formulate a rating. In the second case, we are asked to compare the items and pick the best and worst. In general, whenever we have two sets of different stimuli, we can expect two different cognitive processes converting preferences into responses.

The examples above show different kinds of survey stimuli. But the difference in cognitive processes also occurs when we are analyzing survey data with real-world behavioral data. Even with a realistic looking shelf set, the survey environment is a different stimulus from an actual store. The context of choosing items on a computer screen is different than a real shelf set. While we should expect similarities between the choices, there will likely be some differences in the cognitive processes converting preferences to choices.

Our analysis in turn should account for these different cognitive processes. It is clear that the data stacking/augmentation approach does not account for those differences. That approach assumes a single model that explains all the data, with no adjustments to the parameters. And this can be a problem with the data stacking/augmentation approach. In our third approach, we aim to account for the different cognitive processes. We also allow additional flexibility in the modeling.

2.3 Complete Structural Model/Probabilistic Programming

Our third approach is to build a complete structural model. This structure includes the relationship between the cognitive processes converting unknown preferences to responses to stimuli. This structure is visually described below:

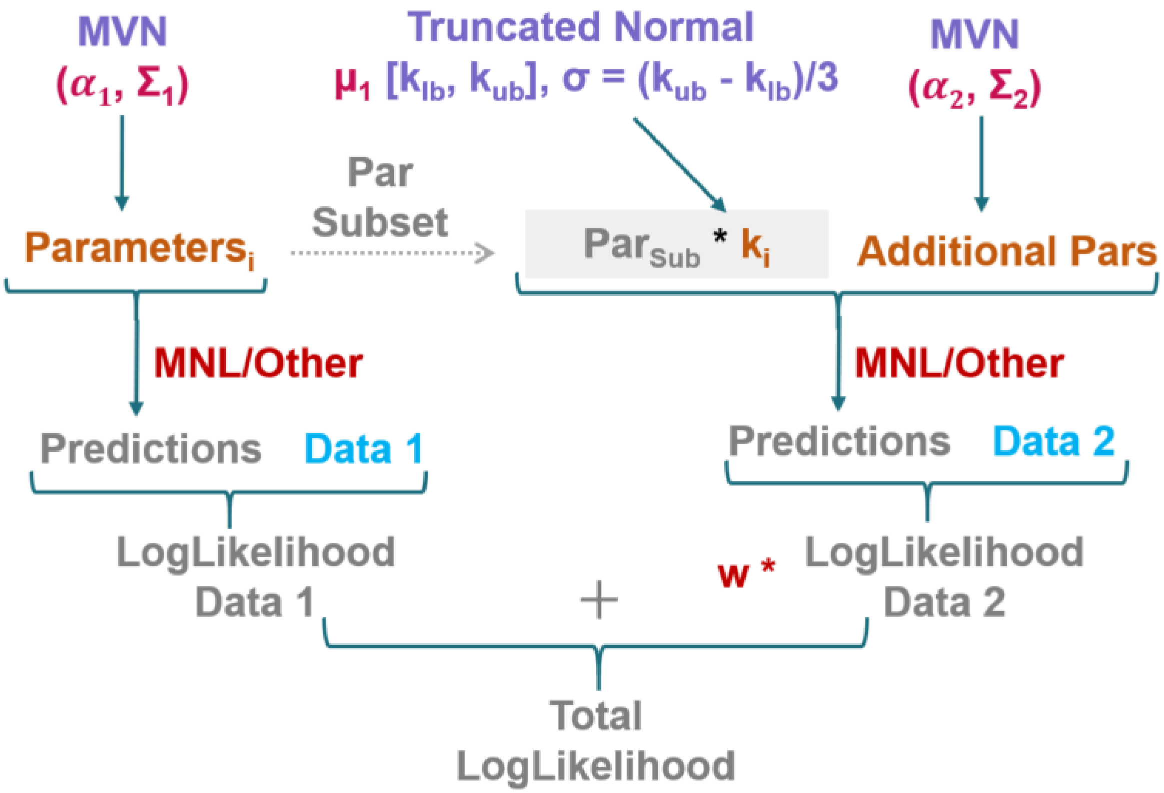


The structure has a left- and right-hand side, corresponding to two different sets of data. As seen at the bottom, we ultimately compute the log-likelihood for each set of data and then sum them. We have also included a weight factor **w** above in case we want to weight the data sources differently. In our case studies we set w = 1, but if one wants to prioritize the predictions for one set of data, the weight can be changed.

On the top left of the chart we have parameters that make predictions for data set 1. The exact nature of the predictive model can be anything, though users of Sawtooth Software are most familiar with multinomial logistic regression (MNL). We then link a subset of those parameters to the right-hand side. This subset might be all the parameters, or a proper subset. In our case we assume this subset of parameters is known ahead of time. We are simply specifying the intersection of the parameters between two different data sets.

The right-hand side of the chart takes the subset of parameters from data set 1 and converts those parameters with a linking function **LinkF**. This linking function is our attempt to model the difference in the cognitive processes from the different stimuli. If the cognitive processes are identical then the linking function would be the identity function. In our case studies, we will use a scaling factor applied to the parameters which are latent utilities. The top right of the chart represents **additional parameters** that are in data set 2 and not in data set 1. (However, our case studies do not actually use any additional parameters.)

The following sections in this paper focus on a specific application of the above structural model. We use a scaling factor for the linkage function and we add priors to the parameters. So, the specific template we will use looks like this:



This more specific version adds multivariate normal priors to parameters for data set 1 and also for those additional (non-overlapping) parameters in data set 2. The scaling factor k_i is respondent-specific and bounded by [lb, ub]. The bounds are defined up-front as data inputs, and we should require that lb >= 0. In our case studies we set the bounds of k to [.1, 3]. We set a truncated normal prior on k, with an unknown mean and a fixed standard deviation defined by the bounds. The standard deviation allows for a broad range of deviations from the mean, at (ub - lb)/3. If one wanted nearly global parameters the deviation could be made very small, like (ub - lb)/50. It would also be possible to allow σ to be an unknown parameter with a hyperprior.

The chart above does not show the hyperpriors, but we use:

Inverse Wishart priors on covariance matrices Σ_1 and Σ_2,
N(0,10) priors on α_1 and α_2, and
Uniform prior on the bounded μ_1.

There is no way to specify such a structure within Sawtooth Software products. The structure in Sawtooth Software's CBC HB is the same that we have specified on the left-hand side above. This specific structure is hard coded within the program with no option to change it. However, we can use more general probabilistic programming languages to specify these structural relationships. BUGS/WinBugs was one of the earliest of these languages and allowed one to specify complex probabilistic structural relationships. Since then, the range of probabilistic programming languages has exploded. Stan is very popular with R users, while PyMC3 is more popular with Python users. In our case studies we use custom R code.

It is worth noting here that our approach is very similar to Dyachenko, Naylor, and Allenby (2013) and to Dyachenko, Reczek, and Allenby (2014). They apply a structural model like the one above to MaxDiff. In those papers MaxDiff is itself treated as a kind of data fusion, with 2 different cognitive processes for two different data sets: the best choice and the worst choice. Horne and Rayner (2013) also suggest that best and worst tasks should be treated as two different data sources with different scales. Dyachenko et al. apply a respondent-level scale factor as the ratio between utilities for the best choice and the worst choice. One key difference from our template is that they estimate a sequential order effect, deriving whether the respondent picked the best or the worst option first. Our template has no such sequential order parameter. However Dyachenko's combination of sequential order effect and scale parameter can be seen as a general linking function. Finally, our approach uses a truncated normal prior on the scale factor, whereas they use a log-normal.

It is also worth noting that Sawtooth Software's ACBC has an option to estimate scale factors for different parts of the ACBC. There are three different parts of an ACBC survey: a screener, a build your own, and a traditional conjoint. These 3 parts can be seen as 3 different data sources. In the ACBC program, one can estimate these 3 parts jointly, either by using simple data augmentation/stacking or by using Otter's Method (Otter 2007), which is a structural model. Otter's structural model uses a global scale parameter (vs. our respondent-specific parameters) and a log-normal (vs. truncated normal) prior for the scale factors. Nonetheless, Otter's method in ACBC is another example of data fusion estimated using a structural model with linkages defined by a scale parameter.

3.0 Case Study 1: MaxDiff with Binary Anchors and Rating Scales

Anchored MaxDiff is one of the most commonly used forms of data fusion. It consists of a set of MaxDiff tasks and a second set of data that we use to anchor the utilities. This anchor data can be obtained in several ways. Louviere suggested a follow-up to the MaxDiff task, asking whether all, some, or none of the options are important. Lattery (2010) proposed a direct binary choice after all the MaxDiff tasks. This binary approach in turn has been extended to asking more granular rating scales. Any of these anchors are cases of data fusion. We will consider both the direct binary and rating scale anchors here, as it is instructive to compare them.

Our case study is an anchored MaxDiff with 23 items. There were 774 respondents, with each respondent completing 15 MaxDiff tasks. Each task showed 5 items, where the respondent selected the best and the worst item. At the end of the 15 MaxDiff tasks, 12 of the items were randomly selected and shown to the respondent in a grid type format. The respondent then rated each of the 12 items on a 5-point Likert scale.

For purposes of testing we created holdout tasks by selecting 3 MaxDiff tasks and 3 rating scale questions. Estimation was based on the remaining 12 MaxDiff tasks and 9 rating scale questions.

We also created a direct binary anchor by using the real data above to create a simulated set of data. We wanted to make the direct binary anchor somewhat different. So, we estimated the MaxDiff utilities (without any anchor) from the initial data. We computed the covariance of the utilities. We then scrambled the variables and drew utilities from the scrambled covariance matrix. We used the same MaxDiff design as the real study. We used all 15 MaxDiff tasks and computed responses from the simulated utilities using Gumbel error. We also added 10 MaxDiff tasks as holdouts (these were newly created).

For the binary anchor, we simulated a threshold for each respondent from N(μ = 1.5, σ = 1). If (MaxDiff utility + Gumbel error) > threshold then the item had a binary value of 1, otherwise it was 0. This gave us simulated binary anchors for each respondent on all 23 items. The mean across all respondents and all 23 binary anchors was 37%. Finally, we randomly selected 10 of the binary anchors for estimation, while the remaining 13 binary anchors were used as holdouts.

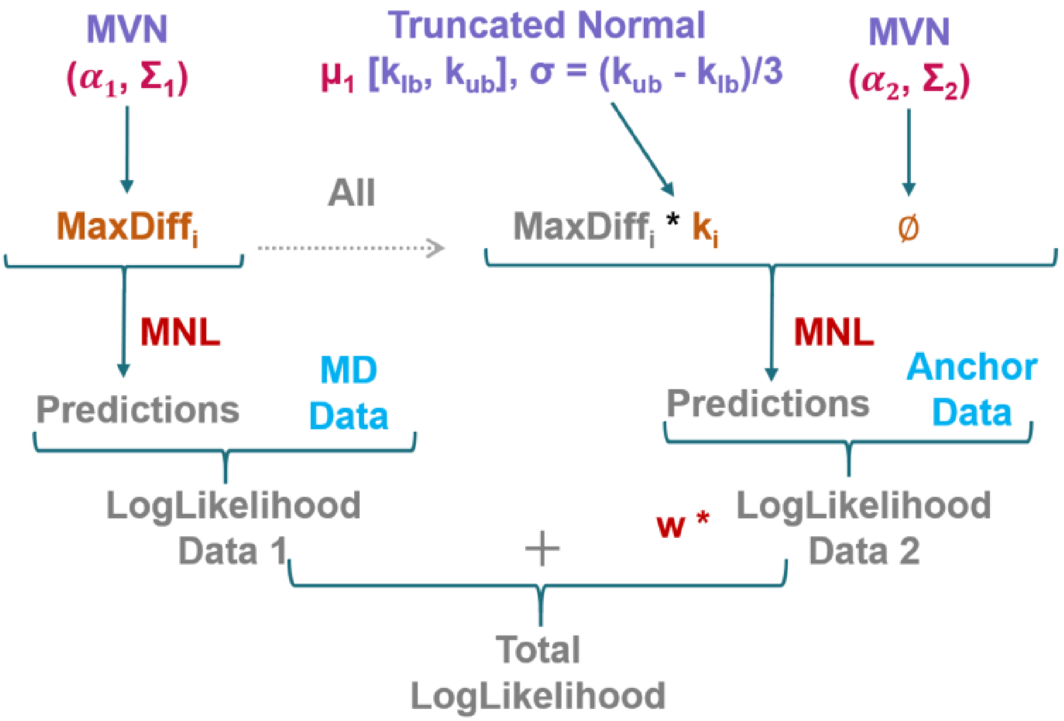
So, we have an initial data set of MaxDiff data anchored by rating scales and a simulated data set with a direct binary anchor. The simulated data set is loosely related to the real data set in that it uses the same design, same number of respondents, same covariance, but with scrambled variables. We estimated both anchored MaxDiffs using 2 different approaches:

1. The simpler data augmentation/stacking method and
2. The structural model with scale factor adjustments.

3.1 Estimation of Direct Binary Anchor

The direct binary method is usually estimated by data augmentation using the method described by Lattery (2010). We create a vector of zeroes as the reference level for each item (no effects or dummy coding). Items chosen beat the 0 vector while those not chosen lose to the 0 vector. We then stack the standard MaxDiff tasks together with the supplemental MaxDiff tasks for winning or losing to the 0 vector.

For the structural model we used the template described in section 2.3, with no additional parameters, so the MVN(α_2,Σ_2) prior does not apply in this case.



This is very similar to the stacked data approach. The MaxDiff utilities on the left-hand side are indicator coded just like we did in the stacking method. The binary anchor tasks are on the right-hand side. Each anchor task has two alternatives with a 0 vector for the anchor. This 0 vector is still the reference level for the MaxDiff utilities. The only difference from the stacked data set is that we have two data files with a respondent-level scale factor adjustment for the binary choices made in the anchor data.

3.1.1 Estimation of Rating Scale Anchor (Augmentation)

Estimating the MaxDiff with a rating scale anchor is more complex. For the data stacking we have to align the two sets of data. This means the two data sources need the same kind of dependent variable. But we have MaxDiff choices (which are 0/1 realizations of probabilities) and ratings (which are Likert scales). One option is to convert the rating scale responses to probabilities. The following method was presented by Lattery at several conferences:

We first define a lower and upper bound for our *predicted ratings*. At first glance it might seem natural to define these bounds as the actual rating scale points. However, this will result in severe underprediction of the end points because we can never quite predict the endpoints, in this case a 1 or a 5. Instead, for a 5-point scale we recommend choosing lower and upper bounds of .5 and 5.5 respectively. This means our predictions will go below 1 and exceed 5. But we can easily infer what those ratings would be. Any predicted rating between .5 and 1.5 is an observed rating of 1, and likewise any predicted rating between 4.5 and 5.5 is a 5. Using the [.5, 5.5] interval, each scale point from 1-5 has a predicted range of 1. Using [1, 5] would mean the lower and upper bound have a range of only .5 while the others would have a range of 1.

Given a lower and upper bound we then convert the ratings to probabilities using the formula:

$$\text{Ratings as Probabilities} = (\text{rating - lb}) / (\text{ub - lb})$$

For a 5-point scale with bounds of [.5, 5.5] the conversions look like this:

Rating	Prob(win)
1	0.1
2	0.3
3	0.5
4	0.7
5	0.9

We can now add the anchoring tasks just as we did with binary anchors. We still have two alternatives in each task, but instead of 0/1 for the choices, we have an allocation probability. If an item is rated a 5, it beats the 0 vector with a probability of 0.9, and loses with a probability of 0.1. The probability of beating the 0 vector is given by the probabilities in the table to the right and the probability of losing to the 0 vector is 1 minus that.

After estimation, we will have a model that makes predictions of probabilities. These probabilities can then be converted back to rating scales using the inverse of our initial conversion. The inverse of “ratings to probabilities” is given by this formula:

$$\text{Probabilities as Ratings} = \text{lb} + (\text{ub} - \text{lb}) \times \text{probability}$$

This will give us ratings in the range of [lb, ub]. Predicted probabilities greater than 0.9 will be converted to ratings between 5 and 5.5, which means we are very confident the respondent gave a rating of 5.

The data augmentation approach must create a single type of dependent variable. So we showed how to convert ratings to probabilities. We then model the combined data set and convert probabilities back to ratings. This conversion to probabilities is not ideal. With a complete structural model, we do not convert ratings to probabilities—we can model the ratings and the MaxDiff choices as two separate sets of data.

3.1.2 Estimation of Rating Scale Anchor (Structural)

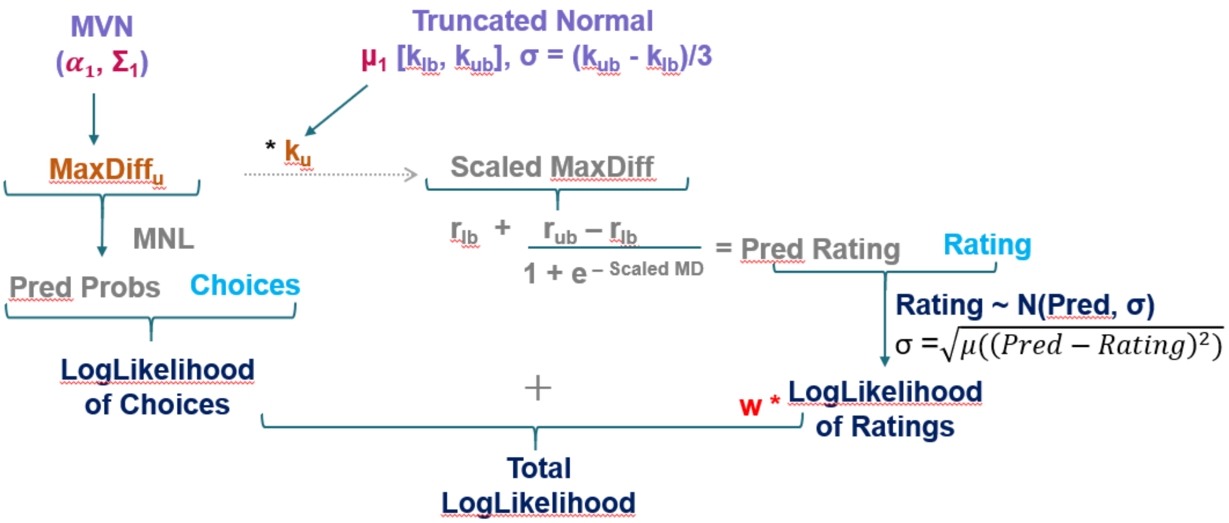
For the structural model we will have two sets of dependent variables. The choices will be predicted from utilities via an MNL model. Like the data augmentation we assume that ratings are continuous. But here we specify the standard log-likelihood version of OLS for continuous data:

$$\text{Predicted Ratings} \sim \text{Normal}(\text{Observed Ratings}, \sigma)$$

Since utilities are in [-∞, ∞] and our ratings are finitely bounded we can use a function that converts numbers in [-∞, ∞] to numbers in bounded intervals [lb, ub]. One such function is an anti-logit with linear transformation to [lb, ub]:

$$\text{lb} + (\text{ub} - \text{lb}) / (1 + e^{-x}), \text{ where x is in } [-\infty, \infty]$$

We still assume that this transformation does not correctly scale the MaxDiff utilities as inputs into the function above. So, for each respondent we rescale them with a respondent-specific factor k_i. The chart below shows the same general structure we had before with the above elements filled in. As before, there are no "additional" parameters on the right-hand side, so they are not shown here.


MVN
(α_1, Σ_1)
Truncated Normal
$\mu_1\ [k_{lb}, k_{ub}], \sigma = (k_{ub} - k_{lb})/3$
$* k_u$
MaxDiff_u
Scaled MaxDiff
MNL
$r_{lb} + \frac{r_{ub} - r_{lb}}{1 + e^{-\text{Scaled MD}}}$ = Pred Rating
Rating
Pred Probs
Choices
Rating ~ N(Pred, σ)
$\sigma = \sqrt{\mu((Pred - Rating)^2)}$
LogLikelihood of Choices
+
w * LogLikelihood of Ratings
Total LogLikelihood

The left-hand side is the same as the binary case, where the utilities are estimated using indicator coding. Adding a constant to the utilities will not change the MaxDiff predictions but it will shift the ratings that get predicted up or down. So, the reference level for the utilities is the ratings, or perhaps the location of the ratings (relatively high or low). Then we link to the right side with a scale factor k. Finally, we convert those scaled utilities to bounded ratings in [lb, ub] and compare those predictions with observed ratings via the log-likelihood version of OLS.

Just as we did with the data augmentation approach, we have to pick lower and upper bounds for our predicted ratings. For our 5-pt scale we pick [.5, 5.5] for the same reasons we discussed previously.

As a side note, it is possible to change the structural model above. One obvious alternative is to code the MaxDiff utilities using effects (or dummy coding), and then estimate a separate threshold parameter on the right-hand side. This creates a separate prior for the threshold, and changes the upper multivariate normal for the MaxDiff utilities. We typically use this approach for dual response None conjoint (with conjoint utilities instead of MaxDiff utilities). But for MaxDiff (with its single attribute), we have found the separate threshold to not predict the rating scale as well.

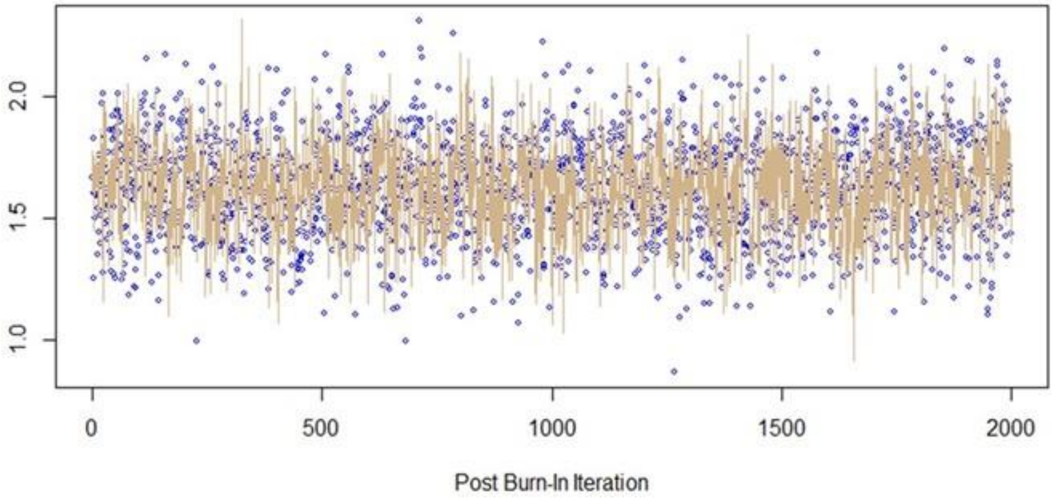
3.2 Scale Factor Results

For all of our estimations we ran two chains and checked for convergence with a Gelman-Rubin test. The Gelman-Rubin results for all parameters were very close to 1 (indicating excellent convergence).

One of the most interesting findings concerned the scale factor parameters. The upper-level mean of the scale factor (μ_1 in the chart above) shows excellent sampling for both the binary and rating scale MaxDiffs.

The traceplot below shows two chains for the binary anchor μ_1, where one chain is represented by lines and a second chain is represented by points:

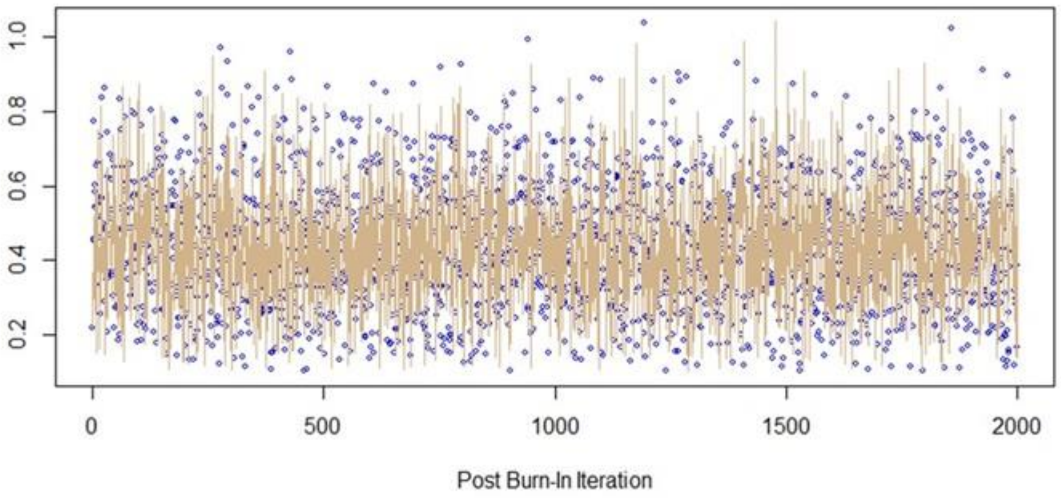
MCMC Sampling of μ_1—Binary Anchor



Both sets of draws for the binary anchor scale factor had a mean of 1.6, with a plausible range. This is about what we would expect going from 5 alternatives in the MaxDiff to 2 alternatives for the Binary anchor. We are effectively increasing the scale from the MaxDiff to the binary anchor. The upper-level mean μ_1 rarely samples below 1.

In contrast, the draws for the scale factor for ratings scales sampled around a mean of .44:

MCMC Sampling of μ_1—Rating Scales

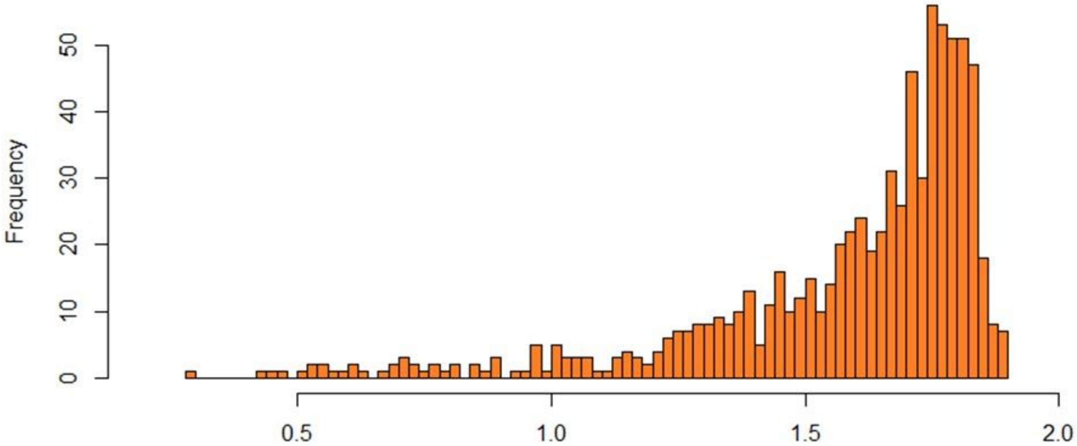


This scale factor μ_1 rarely samples above 1, which means we are effectively decreasing the scale when we move from MaxDiff to rating scales. This is because respondents tend to be somewhat flat in how they use rating scales, while MaxDiff forces them to make tradeoffs.

At the respondent level we see a diversity of scale parameters. For each of the 774 respondents we took their mean scale parameter and show that diversity in a histogram.

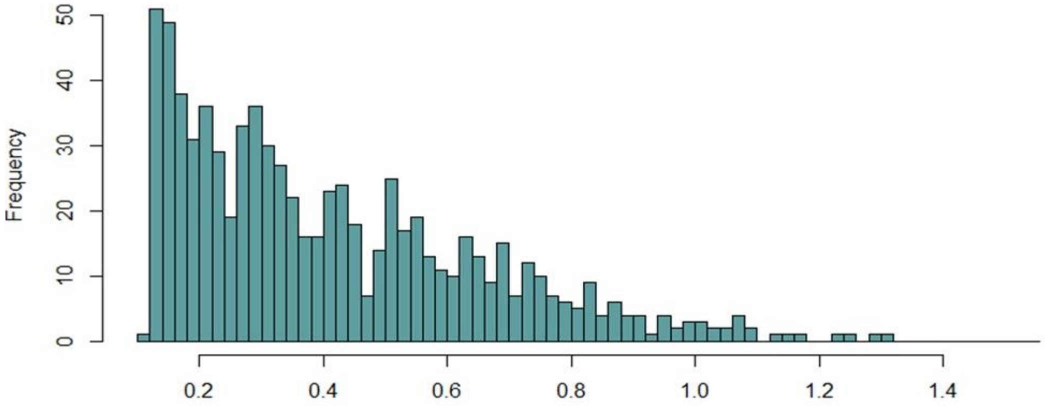
For the binary anchor we see respondents clustering slightly higher than 1.6 (the upper-level mean), and some respondents actually have a mean below 1:

Binary Anchors
Mean Scaling Factor K for Each Respondent



For the rating scale anchor we see respondents clustering lower than .44 (the upper-level mean). In fact many respondents come close to the lower bound we set at .1. These respondents are flatliners, who rated all items the same.

Rating Scale Anchors
Mean Scaling Factor K for Each Respondent



There is also a long tail, with a few respondents having scale parameters greater than 1.

Recall that we sampled k as a truncated normal from [.1, 3], with a σ of range/3 = 2.9/3. This allows respondent-level scale parameters a wide range. If we set σ to a smaller value, we would get far less respondent heterogeneity; as σ goes to 0, we would have a global scaling parameter. Another alternative is to set σ as a parameter that we integrate over with respect to some prior distribution. Both alternatives are very reasonable choices and may be better in some cases than our fixed σ = range / 3.

3.3 Holdout Fit Results

The fit results are very different for binary anchors and rating scale anchors. The model (and theory) suggest the scale factors are significantly different. But for the binary anchor, we did not see this scale factor difference translate to an improvement in fitting holdout tasks. As a result, at this time we do not think scale adjustments are worth the effort for binary anchors. In contrast, the scale adjustments for rating scale anchors gave a significant lift in holdout fit.

A key finding is that the way in which we make predictions is absolutely crucial here. It is common for practitioners to use point estimates of the draws, typically the means. In the standard conjoint model, this often works very well. The respondent-level betas have an upper-level multivariate normal. So, the respondent-level draws are typically normally distributed and the mean point estimates predict nearly as well as draws.

But in our model, the means of the draws for the *right-hand* side of the template (the anchor) are not likely to be nice approximations to the draws. The key difference is that we are multiplying each of the respondent draws by a scale factor k_i. Multiplying each respondent draw $beta_i$ by each respondent draw k_i is very different from multiplying the means of those two. If one must use mean point estimates then one should multiply the two variables and then take the mean of their product, making the anchored utilities the mean of ($beta_i$ * $k_{i.}$), instead of multiplying the means. This works reasonably well, but it is still not as good as using the draws. A final alternative for those who use point estimates is to derive a better point estimate from the posterior, rather than using the mean. We expect to have a future paper on this topic, applying Empirical Bayes to the posterior.

For the *left-hand* side of our model (the MaxDiff) this is less of a problem. These utilities have the same standard multivariate normal prior. So, the respondent-level draws tend to be normally distributed and the mean point estimates work reasonably well. We realize that many practitioners use MaxDiff utilities as substitutes for rating scales and want to use point estimates. These mean point estimates remain a reasonable approximation to the draws.

Binary Anchor

For the binary anchor, we see no difference in how the models fit the MaxDiff data. Here we used the mean point estimates and also added a second simulated data set of binary anchors, to illustrate the stability of the results.

		Augmented/ Stacked	Structural w/Scale K	Augmented/ Stacked	Structural w/Scale K
MaxDiff Mean of Draws	RLH[1] Best	0.701	0.704	0.704	0.706
	RLH Worst	0.700	0.704	0.711	0.711

The binary anchor data is fit slightly better when we use draws. But if we use just point estimates the fit is worse.

		Augmented/ Stacked	Structural w/Scale K	Augmented/ Stacked	Structural w/Scale K
Binary AnchorR LH	Point Estimate 1	0.782	0.735	0.786	0.742
	Point Estimate 2	Above	0.755	Above	0.774
	Draws	0.783	**0.794**	0.788	**0.793**

Here, "point estimate 1" refers to using the mean(beta$_i$) * mean(k$_i$), where i indexes the respondent draws (the simple approach we do *not* recommend). "Point estimate 2" refers to using the means of (beta$_i$ * k$_i$), the preferred approach. Of course, in most cases the practitioner probably does not care about how well they are predicting the binary anchor tasks. In such cases the right-hand side of the model is not used for anything.

Rating Scale Anchor

The rating scale anchor used real responses, not simulated ones. This means respondents might answer the MaxDiff and the rating scale questions inconsistently. A respondent might pick A over B, but rate B more highly. In addition, when we ask rating scale questions we usually do so because we care about making predictions on rating scales. The right-hand side of our model matters, not just for estimation, but also for prediction.

When it comes to predicting the MaxDiff tasks, the holdout fit for data stacking and the structural model are very similar:

[1] RLH is "root likelihood," the geometric mean of the individual likelihoods, a standard measure of model fit.

		Holdout Tasks	
		Augmented/ Stacked	**Structural w/Scale K**
MaxDiff Mean of Draws	RLH Best	0.688	0.663
	RLH Worst	0.705	0.694

The structural model with scale factor fits the MaxDiff data slightly poorer in order to fit the rating scales much better (the numbers shown here are mean squared errors):

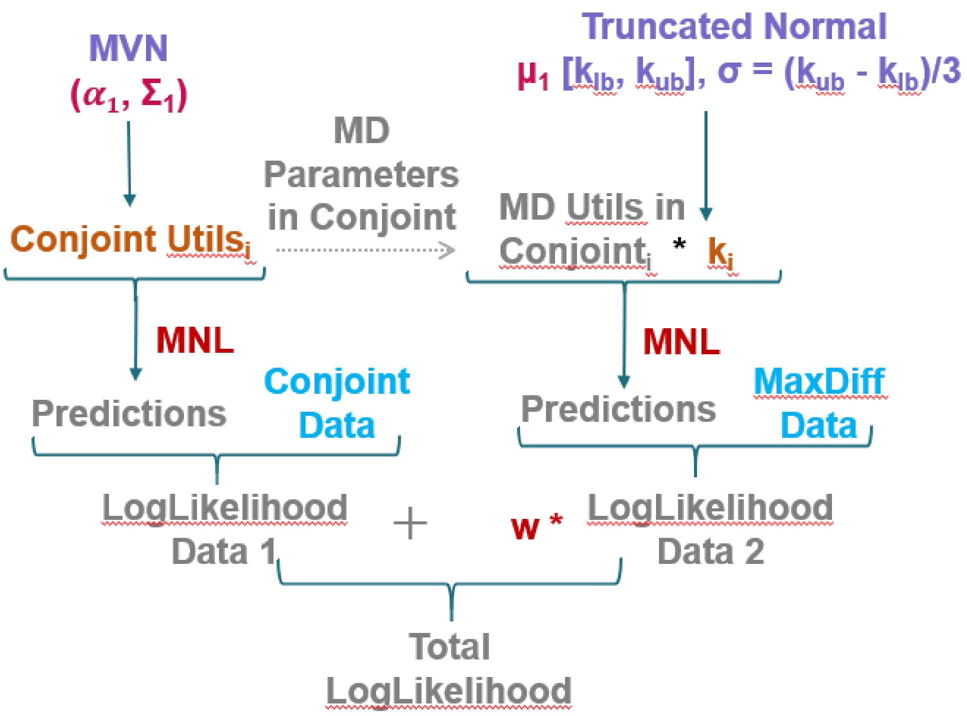
		Holdout Tasks	
		Augmented/ Stacked	**Structural w/Scale K**
5 Pt Rating Scale Mean Squared Error	Point Estimates	1.351	**0.647**
	Draws	1.353	**0.531**

We can see that the scale parameter is very effective at adjusting the utilities to fit the ratings. The draws are still much better than the point estimates, but both are clearly better than the simple stacking of a data augmentation.

4.0 Case Study 2: Conjoint and MaxDiff

We will show another example of the structural model template, this time modeling conjoint and MaxDiff. In this case study, we had 628 respondents. Each respondent completed 9 MaxDiff tasks. There were 25 total items in the MaxDiff, with each task showing 5 of those. In addition, respondents completed a conjoint. The 25 items in the MaxDiff were one of the attributes in the conjoint. Moreover, this attribute only applied to one brand (the client brand). Since one of the key questions was knowing the performance of these levels for the client brand, we supplemented the conjoint with the MaxDiff. Otherwise, we would have very sparse readings on this one attribute from the conjoint.

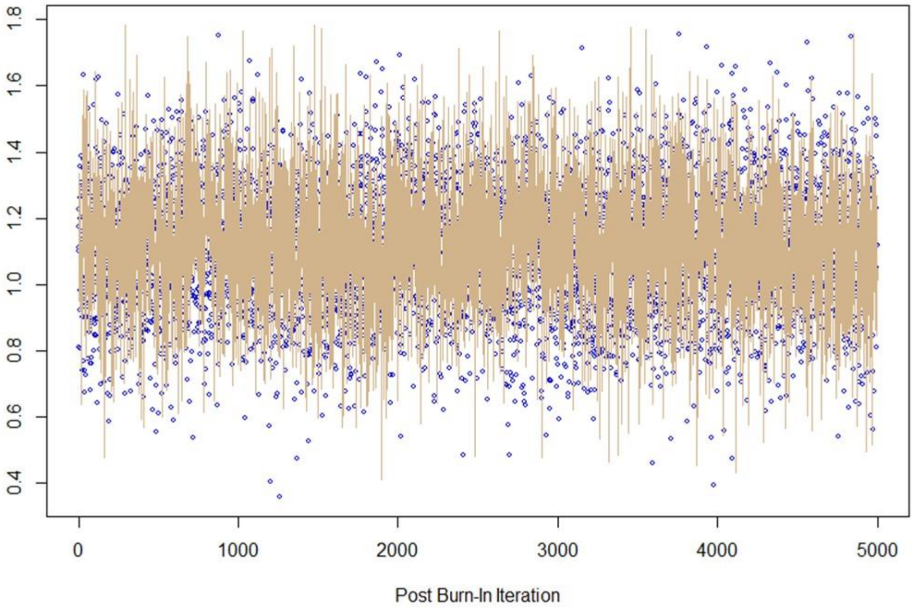
In addition to the MaxDiff attribute, the conjoint measured brand, price, and 2 other attributes. Each respondent completed 9 conjoint tasks as well, each of which showed 6 concepts as choices. The structural chart follows the same pattern as before. In this case, the conjoint utilities and data are on the left-hand-side, with the MaxDiff data on the right-hand side:



One of the differences in this model is that the right-hand side is only a subset of the left-hand-side parameters, namely just the one attribute corresponding with the MaxDiff and its 25 levels. Here we used the last MaxDiff level as a reference, so there are 24 parameters that correspond. In other studies we have tested a subset of the MaxDiff items in the conjoint, choosing those items thought to be representative. In some cases we have used on-the-fly MaxDiff estimation to select the items to include in the conjoint. Here all 25 items appeared in the conjoint, which we think makes a more useful research test.

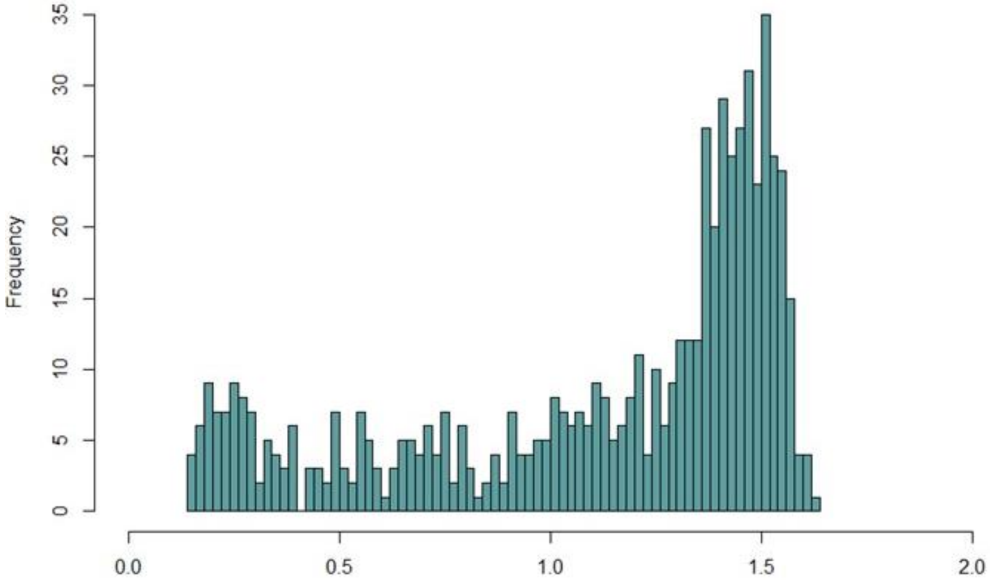
We again ran two chains and found excellent convergence with the Gelman-Rubin test. The traceplot below is (as before) for the upper-level mean of the scale factor (μ_1 in the chart above). We expected the scale factor of the MaxDiff utilities to be greater than their corresponding utilities in the conjoint. Indeed, we found the upper-level mean hovers around 1.2, although we expected it to be higher still.

MCMC Sampling of μ_1



The mean of the scale factor draws across respondents is also interesting. Many of the respondents are near 1.5, but quite a few respondents show scale factors less than 1.

Mean Scaling Factor K for Each Respondent



Even though these diagnostics are not quite in line with our initial expectations, the structural model still outperformed the others in predicting holdouts.

To create holdouts, we removed one conjoint task and one MaxDiff task for each respondent. We estimated the models using the remaining tasks (8 conjoint, 8 MaxDiff) and predicted the holdouts. We then repeated this 2 more times for a total of 3 holdouts for the conjoint and 3 for the MaxDiff.

One model we ran used just the conjoint data. For this model we ignored the MaxDiff data completely. We thought this should provide a baseline for how our model should predict the conjoint, and how well those predictions align with the MaxDiff tasks.

	Holdout RLH		
	Conjoint-Only Model	**Augmented/ Stacked**	**Structural w/Scale K**
Conjoint	0.429	0.384	**0.441**
MaxDiff Best	0.245	0.393	**0.410**
MaxDiff Worst	0.216	0.333	**0.366**

The conjoint-only model fits the conjoint data fairly well with an RLH of .429. But it fits the MaxDiff data much more poorly than the other models. Bear in mind that an RLH of .2 is effectively random with 5 choices. This illustrates our point that one should not trust the data gods to align the data scales for you.

The stacked and structural models both fit the MaxDiff data much better, with the structural model consistently best. We are not fans of hit rates, because they are very unstable. But they do eliminate any bias in scale, so we computed the hit rates here as well:

	Holdout Hit Rates		
	Conjoint-Only Model	**Augmented/ Stacked**	**Structural w/Scale K**
Conjoint	51.6%	51.6%	52.4%
MaxDiff Best	**29.8%**	91.7%	88.3%
MaxDiff Worst	**24.8%**	92.0%	90.2%

We still see that the conjoint-only model still fits the MaxDiff tasks just slightly better than random (which is 20%). The hit rates of the other two methods are much higher. In fact, we worry that both are too high at 88%+. While the stacked approach has a higher hit rate for MaxDiff, we do not think that is relevant.

All three approaches fit the conjoint holdouts about equally well. The lack of alignment between conjoint-only and MaxDiff initially led us to speculate that a joint model of MaxDiff and Conjoint would seriously compromise the conjoint. But the holdout RLH and hit rates show that is not the case. We get results that are much more consistent with MaxDiff, and as good (or even a bit better) for the conjoint.

5.0 Summary

We described three general approaches to data fusion:

1. Two-Stage Linkage
2. Data Augmentation/Stacking
3. Complete Structural Model/Probabilistic Programming

We should also note that these three approaches are not mutually exclusive. It is possible to apply data augmentation/stacking to some of the data and fit that with a larger structural model. We could even potentially fit that with a two-stage linkage. We think of these as three general approaches. Our general preference is for a complete structural model. The ideal of Probabilistic Programming is to fit any kind of structural model you can describe, though we are not there yet.

We have emphasized that when we do data fusion we must take into account different stimuli and cognitive processes. In general, whenever we have two sets of different stimuli, we can expect two different cognitive processes converting preferences into responses. We focused on thinking of those cognitive processes as utility mappings with a simple scale conversion linking utilities between two cognitive processes. This scaling linkage has additional precedent in the types of data fusion studies here.

Another key finding is that the way in which we make predictions matters. It is common for practitioners to use point estimates of the draws, typically the means. In the standard conjoint model, this usually works very well. The respondent-level betas have an upper-level multivariate normal prior. So, the respondent-level draws are typically normally distributed and the mean point estimates predict nearly as well as draws. But for more complex structural models the mean point estimates are no longer accurate approximations. So, for more complex structural models we recommend using draws, as most of the Bayesian world does. For those who must use point estimates we recommend approaches other than the simple mean, another topic we plan to discuss in a future paper.

We showed a basic structural template and how that could be applied to many cases. We specifically discussed:

Data Set 1	Data Set 2
MaxDiff	Anchor Question
MaxDiff	Purchase Intent Ratings
Conjoint	MaxDiff

For a simple anchored MaxDiff we did not see much benefit in using the structural model. But other data sets may show the need for such a structural model. Our data fusion model was especially valuable in the context of MaxDiff plus rating scales.

The details in the models we described can be modified. For instance, we treated the purchase intent ratings as continuous, but they could also be treated as ordinal, perhaps with better results.

Although we only illustrated the 3 models above, it should be clear that the template can also be used for cases like:

Data Set 1	Data Set 2
Conjoint	Buy or Not
Conjoint	Purchase Intent Ratings
Conjoint	Ratings of Levels

And, of course, our template is just one basic framework that applies to some cases. There are also many other structures one can use in data fusion.

While we did not discuss it here, some of our key data fusion work involves survey and real-world data. In that case, we cannot link respondents, and instead we might link parameters at the upper level. For instance, we might hypothesize that the survey respondents and real-world respondents have the same upper-level covariance structure, perhaps adjusted by a scale factor. We look forward to sharing this work at a later time, and in the meantime have focused on applications that might be more commonly found among marketing research practitioners.

Kevin Lattery

REFERENCES

Fusion via Data Augmentation

Bahna & Chapman (2018). "Constructed, Augmented MaxDiff." *2018 Sawtooth Software Conference Proceedings*, 1–12.

Chrzan & Yardley (2009). "Tournament Augmented Choice based Conjoint." *2009 Sawtooth Software Conference Proceedings*, 163–170.

Fuller, Madden, & Smith (2013). "Augmenting Discrete Choice Data—A Q-Sort Case Study." *2013 Sawtooth Software Conference Proceedings*, 97–104.

Hendrix & Drucker (2007). "Alternative Approaches to MaxDiff with Large Sets of Disparate Items—Augmented and Tailored MaxDiff." *2007 Sawtooth Software Conference Proceedings*, 169–188.

Johnson & Orme (2007). "A New Approach to Adaptive CBC." *2007 Sawtooth Software Conference Proceedings*, 85–110.

Jones & Yeh (2013). "MaxDiff Augmentation: Effort vs Impact." *2013 Sawtooth Software Conference Proceedings*, 105–114.

Lattery, Kevin (2010). "Anchoring Maximum Difference Scaling Across a Threshold." *2010 Sawtooth Software Conference Proceedings*, 91–106.

Lattery, Kevin (2009). "Coupling Stated Preferences with Conjoint Tasks to Better Estimate Individual Level Utilities." *2009 Sawtooth Software Conference Proceedings*, 171–184.

Fusion via Structure

Dyachenko, Naylor, Allenby (2013). “The Ballad of Best and Worst.” *2013 Sawtooth Software Conference Proceedings*, 357–366. More technical document in *Marketing Science* (Nov 2014).

Dyachenko, Reczek, Allenby (2014). “Models of Sequential Evaluation in Best-Worst Choice Tasks.” *Marketing Science*, 33:6 (November–December 2014), 828–848.

Otter, Thomas(2007). “HB-Analysis for Multi-Format Adaptive CBC.” *2007 Sawtooth Software Conference Proceedings*, 111–126.

Other

Horne & Rayner (2013). “Does the Analysis of MaxDiff Data Require Separate Scaling Factors?” *2013 Sawtooth Software Conference Proceedings*, 331–340.

Karty, Kevin (2010). “Integrating Self Stated Data into Choice Models without Bias.” 2010 AMA Advanced Research Techniques (ART) Forum.

SEGMENTING CHOICE AND NON-CHOICE DATA SIMULTANEOUSLY: PART DEUX

THOMAS C. EAGLE
EAGLE ANALYTICS OF CALIFORNIA, INC.
JAY MAGIDSON
STATISTICAL INNOVATIONS, INC.

BACKGROUND

There are two primary motivations for this paper. First, we wish to reconsider advice from the previous Sawtooth Software Conference paper by Eagle (2013), given to the many practitioners segmenting raw or rescaled HB-derived choice model parameters. Second, there are two new advances in latent class (LC) modeling we wish to introduce: the ability to apply a scale factor to continuous and count variables; and the ability to weight the impact of one or more variables on segment solutions. In this paper, we revisit the advice that practitioners should not segment respondents using hierarchical Bayesian (HB) derived parameters. Would the inclusion of scale classes in a LC model enable practitioners to obtain meaningful segments based on HB derived parameters?

Given general agreement that clustering on HB utilities often yields results that are difficult to interpret and not reproduceable by more theoretically appropriate approaches, why do HB practitioners continue to segment using HB choice model parameters, scaled or not? The answer is that they do so out of convenience, since they have these parameters at hand. It is also easy to combine choice data results (i.e., the HB utilities) with non-choice data such as attitudes, behaviors, and other respondent data, to produce a single combined choice and non-choice rectangular data file for segmentation. In contrast, constructing a proper data file within the LC framework consisting of responses to both choice and non-choice data is more difficult (see Appendix C).

In the past, Sawtooth Software recommended rescaling the HB parameters to reduce the impact of scale differences across respondents prior to segmentation. However, Eagle (2013) demonstrated that segments are heavily influenced by the magnitude and type of rescaling. The key conclusion of the 2013 paper was that it is bad practice to segment based on HB choice model parameters.

More recently an Advanced Research Techniques Forum presentation by Lee and Brazell (2019) suggested ANY multivariate analysis conducted on derived HB choice model parameters is suspect. As a note, Sawtooth Software no longer recommends segmenting derived HB choice model parameters, rescaled or not. Despite these and other warnings, HB practitioners continue to segment using HB choice model parameters. In the current (2019) Sawtooth Software conference itself, several papers and tutorials included segmentation of HB choice model parameters.

Coincidentally, the authors of a paper on volumetric models (Eagle et al., 2018) made a request to the Latent GOLD developers to extend Scale-Adjusted Latent Class (SALC) models to apply to *count* variables (e.g., the Poisson, the negative binomial, and zero inflated models). As a result, version 6.0 of Latent GOLD extends SALC modeling to more traditional forms of LC cluster modeling based not only on categorical and count variables, but also continuous variables. It was then realized that the latter extension would allow SALC modeling to be conducted on HB parameters treated as continuous variables.

In this paper, we begin with an examination of a Best-Worst (i.e., "MaxDiff") case study in which we compare segmentation solutions derived using a LC *choice* model framework to analyze choice responses directly, with that of a LC *cluster* analysis to segment on derived HB utility parameters. We examine both methods first without taking into account scale heterogeneity, and then again using SALC models which account for scale heterogeneity explicitly.

We follow this by analyzing a more complex data set to examine the impact of including non-choice data with Best-Worst choice data in the segmentation. As described above, we compare two methods: an LC model that analyzes choice and non-choice responses simultaneously, and LC clustering of HB parameters associated with both the choice and non-choice data. Again, we examine the two methods first without taking into account scale heterogeneity, and then again using SALC models to account for scale heterogeneity in the choice variables.

As we investigated the more complex data application, we noticed that the resulting segments obtained from both methods appeared to be influenced more heavily by the choice, as opposed to the non-choice, variables. This led to the development of the second advance in Latent GOLD 6.0—a variable weighting capability within the latent class framework that enables the researcher to differentially weight sets of variables in the construction of segments. Results using this advance to weight the choice variables less than the non-choice variables are examined.

Brief Introduction to Scale Confounds and the SALC Model

The goal in LC choice modeling is to identify homogeneous segments, each differing in respondent preferences. However, *standard* LC choice models derive segments that differ not necessarily in preferences, but in Utilities = Preferences * Scale. Respondents who are unsure of their preferences have "low scale" values and are less consistent in their responses (for pure random responders, scale = 0). As a group, these less consistent respondents have similar utilities (for this group all utilities are small) and thus tend to be clustered together by latent class analysis into their own "low scale" segment. However, unlike the other segments, each of which is homogeneous in their preferences, the "low scale" segment often contains a *heterogeneous* group of respondents with different (but somewhat weak) preferences.

By explicitly separating scale and preference parameters, SALC modeling allows all respondents to be assigned to their most likely *preference* class irrespective of how certain they are regarding their preferences—the utilities for all respondents within a given preference class being proportional, the respondent's scale factor serving as the constant of proportionality. Standard LC choice modeling can't do this because no separate scale parameter is included in these log-linear models. As such, the resulting utility or part-worth parameters inextricably confound preference and scale.

In contrast to the LC choice model, the SALC model re-expresses utility in a log-*bilinear* form,[1] as the product of separate *preference* and *scale* parameters, and utilizes two distinctly different latent variables—the first is specified to be the source of the *preference* heterogeneity, affecting only the preference parameters, and the second, as source of the *scale* heterogeneity, affecting only the scale parameters. The first latent variable is discrete, and since it is defined solely by the *preference* parameters, its categories can rightfully be interpreted as *preference classes*.

The second latent variable in the SALC model, the scale, can be specified as either continuous or discrete. In this paper, when specified as *discrete*, we refer to its categories as *scale classes* (e.g., "low scale," "medium scale," and "high scale" respondents in the case of S=3 scale classes). Thus, each respondent belongs to one of K preference classes (K segments differing in their preferences) and also belongs to one of S scale classes, each scale class having its own scale parameter. That is, each respondent is simultaneously a member of one preference segment and one scale class. More detail on SALC models is provided in Appendix A (see also Groothuis-Oudshoorn et al., 2018).

Example 1: Analysis of Best-Worst Data from Australian Health Reform Study

For our initial set of analyses, we use Best-Worst data from the Australian Health Reform Study (Louviere and Flynn, 2010), to compare LC segments obtained with and without scale adjustment. The authors thank Terry Flynn for providing the data for this research. Flynn hypothesized three preference segments:

"In health economics you usually find people separate out into 3 policy-relevant classes

- those who value equity
- those who value efficiency/value for money
- those who value investment in future health" . . . Terry Flynn

Our analyses are confirmatory and attempt to confirm Flynn's hypothesis that there are three segments with these distinctly different preferences.

The different types of segmentation models can be grouped into four categories as shown in Table 1.

Table 1. Four types of Latent Class Segmentation Models

Scale adjustment?	(A) Analysis of Choice responses	(B) LC Clustering of HB Utilities Derived from Choice Responses
NO—LC model	**(A1) LC Choice model**	**(B1) LC Cluster model**
YES—SALC model	**(A2) SALC Choice model**	**(B2) SALC Cluster model**

[1] See Appendix A for details on the log-bilinear form.

Specifically, our research goals are as follows:

A. **Segmentation based on LC analyses of Best-Worst responses**. Comparison of LC segments obtained with and without scale adjustment (A1 vs. A2). We compare (A1) segments obtained *without* scale adjustment using the LC choice model to (A2) segments obtained *with* scale adjustment using the log-bilinear SALC choice model. In particular, we wish to determine whether the A1 segments show any evidence of scale confounds, and if so, whether such confounds are removed (in A2) by the SALC model.
B. **"Tandem" approach to Segmentation—LC clustering of HB utilities derived from Best-Worst responses.** Comparison of LC segments obtained with and without scale adjustment, but now using the LC *cluster* model (B1) and SALC *cluster* model (B2) to cluster respondents based on their (zero-centered) HB-derived parameters. We then evaluate the resulting segmentations to determine 1) how the LC clustering without scale adjustment compares to LC segmentation of Best-Worst responses obtained without scale adjustment (B1 vs. A1), and 2) how the LC clustering based on the SALC model compares with the best practice result obtained by applying the SALC model directly to analyze the Best-Worst choice responses (B2 vs. A2).

The HB utilities used in the B-type analyses were estimated by Tom Eagle using Sawtooth Software's standalone CBC-HB program with the default setting for identification (zero-centering).[2]

The 15 health principles (items) evaluated by the N= 204 voting age citizens in the Best-Worst experiment are listed in Table 2. For further details of the data and survey design, see Louviere and Flynn (2010).

Table 2. List of the 15 principles evaluated as part of the Best-Worst experiment.

1	People & family centred
2	Equity
3	Shared responsibility
4	Promoting wellness & strengthening prevention
5	Comprehensiveness
6	Value for money
7	Providing for future generations
8	Recognise social & environmental influences shape our health
9	Taking the long-term view
10	Quality & safety
11	Transparency & accountability
12	Public voice & community engagement
13	A respectful, ethical system
14	Responsible spending
15	A culture of reflective improvement & innovation

[2] When SALC clustering is applied to HB utilities (B2 approach), the resulting segments depend on the criteria used to identify the utilities (e.g., zero-centering vs. zero-referencing). Lyon (2020) shows that use of zero-referencing yields segmentations which are very different depending on which item is chosen as the reference. Our decision to use zero-centering (rather than zero-referencing) to identify the HB utilities was intentional, and we caution against blind use of a single item as reference when clustering on HB utilities, especially when applying the SALC cluster model. In contrast, segments obtained from SALC *choice* modeling (A2 approach) do not depend on the criteria (coding) used to identify the parameters (e.g., effect coding vs. dummy coding).

Segmentation on Observed Choices (A-type Choice Models) with and without Scale Adjustment

Table 3 shows the utility parameter estimates obtained for each of the classes in the 3-class choice model (the A1 model of Table 1). Classes 1 and 3 provide evidence for two of Flynn's posited segments—those who value "Value for Money" (Class 1 in Table 3) and those who value "investment in future health" (Class 3 in Table 3). However, the low magnitude of the utility parameters for Class 2 suggests that Class 2 primarily captures "low scale" persons, consisting of 34% of respondents. This belief is reinforced by the bottom row of Table 3, which shows that the standard deviation for the Class 2 parameters is much smaller (0.31) than those for the other classes (1.08 and 0.96).

Table 3: Results for 3-Class Choice Model (A1 Model with 3 Classes)

	Utility estimates		
	3-class LC choice model		
Item	Class 1	Class 2	Class 3
A culture of reflective improvement & innovation	-1.64	-0.48	-0.49
A respectful, ethical system	-0.23	0.34	0.51
Comprehensiveness	-0.24	0.27	-1.08
Equity	-0.11	**0.52**	-1.55
People & family centered	0.49	-0.16	**1.60**
Promoting wellness & strengthening prevention	0.28	0.22	**1.32**
Providing for future generations	0.02	0.10	**0.99**
Public voice & community engagement	-1.72	-0.33	-0.54
Quality & safety	**2.14**	**0.35**	**0.92**
Recognize social/environ influences shape health	-1.14	-0.13	0.55
Responsible spending	**0.92**	0.08	-0.32
Shared responsibility	-0.47	-0.49	-0.81
Taking the long-term view	-0.11	-0.21	0.22
Transparency & accountability	0.02	-0.04	-0.24
Value for money	**1.80**	-0.04	-1.08
Class size	0.35	0.34	0.31
Standard deviation	1.08	**0.31**	0.96

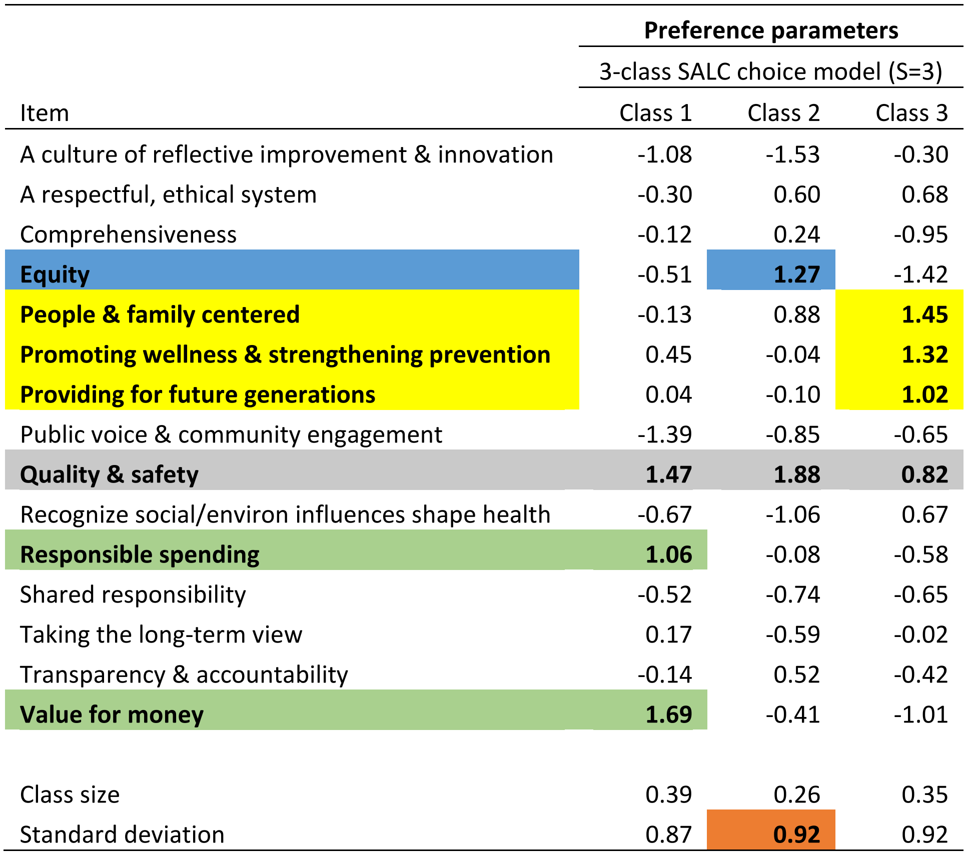
In comparison, results from the 4-class A1 model (Table 4) provide support for the existence of all three of Flynn's segments (Class 4 prefers Equity), but the largest of the four classes (Class 1 which contains 38% of respondents) again appears to be a "low scale" class, now containing 38% of respondents.

Table 4: Results for 4-Class Choice Model (Model A1 with 4 Classes)

	Utility parameters			
	4-class LC choice model			
	Class 1	Class 2	Class 3	Class 4
A culture of reflective improvement & innovation	-0.37	-1.42	-0.42	-2.02
A respectful, ethical system	0.17	-0.38	0.88	0.87
Comprehensiveness	-0.05	-0.13	-1.34	-0.13
Equity	-0.03	-0.66	-1.97	**1.66**
People & family centered	0.16	0.06	**2.05**	1.04
Promoting wellness & strengthening prevention	0.37	0.51	**1.77**	-0.11
Providing for future generations	0.37	0.04	**1.22**	-0.24
Public voice & community engagement	-0.35	-1.92	-0.55	-0.91
Quality & safety	0.07	**2.00**	**1.19**	**2.70**
Recognize social/environ influences shape health	0.10	-0.99	0.90	-1.50
Responsible spending	0.18	**1.21**	-0.88	-0.09
Shared responsibility	-0.57	-0.48	-0.59	-0.88
Taking the long-term view	0.02	0.12	-0.06	-0.61
Transparency & accountability	-0.10	-0.13	-0.69	0.87
Value for money	0.02	**2.18**	-1.51	-0.66
Class size	0.38	0.28	0.19	0.15
Standard deviation	**0.26**	1.10	1.21	1.20

In contrast, the 3-class SALC (A2) model (Table 5) provides strong support for Flynn's three segments, and since the standard deviation for the Equity class (class 2) is similar to the other classes, SALC appears to remove the scale confound evident in the 3-class LC model (Table 3)—the magnitude of the Equity Preference parameter is relatively high and the standard deviation of the class 2 parameters is 0.92, comparable to those of the other classes.

Table 5: Results for 3-class SALC Choice Model (A2; K=3 Preference Classes, S=3 Scale Classes)



	Preference parameters		
	3-class SALC choice model (S=3)		
Item	Class 1	Class 2	Class 3
A culture of reflective improvement & innovation	-1.08	-1.53	-0.30
A respectful, ethical system	-0.30	0.60	0.68
Comprehensiveness	-0.12	0.24	-0.95
Equity	-0.51	**1.27**	-1.42
People & family centered	-0.13	0.88	**1.45**
Promoting wellness & strengthening prevention	0.45	-0.04	**1.32**
Providing for future generations	0.04	-0.10	**1.02**
Public voice & community engagement	-1.39	-0.85	-0.65
Quality & safety	**1.47**	**1.88**	**0.82**
Recognize social/environ influences shape health	-0.67	-1.06	0.67
Responsible spending	**1.06**	-0.08	-0.58
Shared responsibility	-0.52	-0.74	-0.65
Taking the long-term view	0.17	-0.59	-0.02
Transparency & accountability	-0.14	0.52	-0.42
Value for money	**1.69**	-0.41	-1.01
Class size	0.39	0.26	0.35
Standard deviation	0.87	**0.92**	0.92

Table 6 provides an expanded view of the preference parameters in the 3-class SALC model, breaking out the separate parameters for each *scale class* (high, medium, and low) within each *preference class*. For each preference class, the preference parameters displayed in Table 5 are computed as weighted averages across the three scale classes of the parameters displayed in Table 6.

Table 6: Expanded View of Results for 3-Class SALC Choice Model (A2, K=3, S=3)

	Preference class 1			Preference class 2			Preference class 3		
	by scale class			by scale class			by scale class		
Item	High	Med	Low	High	Med	Low	High	Med	Low
A culture of reflective improvement & innovation	-2.17	-1.15	-0.29	-3.06	-1.63	-0.42	-0.61	-0.32	-0.08
A respectful, ethical system	-0.60	-0.32	-0.08	**1.21**	**0.64**	**0.16**	1.37	0.73	0.19
Comprehensiveness	-0.24	-0.13	-0.03	0.48	0.25	0.06	-1.91	-1.01	-0.26
Equity	-1.03	-0.55	-0.14	2.56	1.36	0.35	-2.86	-1.52	-0.39
People & family centered	-0.27	-0.14	-0.04	1.77	0.94	0.24	**2.92**	**1.55**	**0.40**
Promoting wellness & strengthening prevention	0.90	0.48	0.12	-0.08	-0.04	-0.01	**2.64**	**1.40**	**0.36**
Providing for future generations	0.08	0.04	0.01	-0.21	-0.11	-0.03	**2.04**	**1.09**	**0.28**
Public voice & community engagement	-2.79	-1.48	-0.38	-1.71	-0.91	-0.23	-1.30	-0.69	-0.18
Quality & safety	**2.95**	**1.57**	**0.40**	**3.78**	**2.01**	**0.51**	**1.64**	**0.87**	**0.22**
Recognize social/environ influences shape health	-1.35	-0.72	-0.18	-2.13	-1.13	-0.29	1.35	0.72	0.18
Responsible spending	**2.12**	**1.13**	**0.29**	-0.17	-0.09	-0.02	-1.15	-0.61	-0.16
Shared responsibility	-1.05	-0.56	-0.14	-1.49	-0.79	-0.20	-1.30	-0.69	-0.18
Taking the long-term view	0.34	0.18	0.05	-1.18	-0.63	-0.16	0.05	0.02	0.01
Transparency & accountability	-0.27	-0.14	-0.04	1.05	0.56	0.14	-0.85	-0.45	-0.12
Value for money	**3.39**	**1.80**	**0.46**	-0.82	-0.43	-0.11	-2.02	-1.08	-0.27
Class size	0.08	0.19	0.12	0.05	0.13	0.08	0.07	0.17	0.11
Standard deviation	1.69	0.90	0.23	1.79	0.95	0.24	1.78	0.95	0.24

Since the parameters for the three scale classes within each preference class are proportional to each other, all respondents in a given preference class are homogeneous with respect to their preferences. For example, Table 6 shows that the preference parameter for “Value for Money” is highest for all three scale classes within Preference class 1 (see the highlighted values 3.39, 1.80, 0.46), which in each case is the highest among the 15 items evaluated in the corresponding column. Thus, despite reflecting different amounts of uncertainty in their choices, all respondents within Preference class 1 prefer “Value for Money.”

Within each preference class, the ratio of the preference parameters represent *relative scale factors*. For example, the scale factor for the “medium scale” class in Preference class 1 (relative to the “high scale” class) = 1.80/3.39 = .53, while the corresponding scale factor for the “low scale” class is only 0.47/3.39 = .14 times as large as the “high scale: class. (The relative scale factors can also be computed as the ratio of the corresponding Standard deviations: 0.90/1.69 = .53 and 0.23/1.69 = .14).

More detailed interpretation of these results and related statistics are provided in Appendix A. Those results include:

- The 3-class SALC model (model A2 with K=3) fits the data better than either the 3-class or 4-class LC choice models (model A1 with K=3 or K=4).
- The fit of the SALC model is best with 3 scale classes.

- The fit of the SALC model with 3 scale classes is better than that of a SALC model that treats scale as continuous.

Segmentation Based on HB Utilities (B-type Cluster Models) with and without Scale Adjustment

Summary statistics for the HB utilities (Table 7) show that the three largest standard deviations (2.4, 2.4, and 2.1) are associated with the three segments posited by Flynn. Thus, we might expect LC clustering of the individual-level HB utilities would be supportive of Flynn's hypothesis.

Table 7: Descriptive Statistics for the HB Utilities

All Items	Mean	Std. Dev.
A culture of reflective improvement & innovation	-1.4	1.7
A respectful, ethical system	0.3	1.6
Comprehensiveness	-0.5	1.6
Equity	-0.5	2.4
People & family centered	0.9	2.1
Promoting wellness & strengthening prevention	0.9	1.9
Providing for future generations	0.5	1.5
Public voice & community engagement	-1.4	1.8
Quality & safety	1.9	1.9
Recognize social/environ influences shape health	-0.5	1.9
Responsible spending	0.4	1.7
Shared responsibility	-0.9	1.4
Taking the long-term view	0.0	1.5
Transparency & accountability	0.0	1.5
Value for money	0.4	2.4

Items with Std. Dev. > 2	Std. Dev.
Equity	2.4
Value for Money	2.4
People & Family Centered	2.1

Tables 8 and 9 present the parameter estimates from the 3- and 4-class LC cluster models developed using these HB utilities (B1 models), and Table 10 presents results from the 3-class SALC cluster model (B2 model with K=3 preference classes, and S=3 scale classes).

Table 8: Results from 3-Class Cluster Analysis of the Derived HB Utilities (B1, 3 Classes)

Items	Value for the Money	Future	Equity & Value for the Money
A culture of reflective improvement & innovation	0.52	1.09	-1.61
A respectful, ethical system	-0.58	0.63	-0.05
Comprehensiveness	0.15	-0.89	0.75
Equity	-0.11	-1.11	**1.22**
People & family centered	-0.49	**1.11**	-0.62
Promoting wellness & strengthening prevention	-0.03	**1.20**	-1.16
Providing for future generations	0.13	**0.69**	-0.81
Public voice & community engagement	0.08	0.63	-0.71
Quality & safety	**-1.14**	**-0.59**	**1.74**
Recognize social/environ influences shape health	-0.01	**1.93**	-1.92
Responsible spending	**0.68**	-1.69	**1.00**
Shared responsibility	0.02	0.04	-0.06
Taking the long-term view	0.47	-0.10	-0.37
Transparency & accountability	-0.41	-0.69	1.09
Value for money	**0.74**	-2.25	**1.51**
Class size	0.49	0.23	0.27
Standard deviation	**0.49**	1.14	1.12

Table 9: Results from 4-Class Cluster Analysis of the Derived HB Utilities (B1, 4 Classes)

Items	Low Scale	Value for Money	Equity	Future
A culture of reflective improvement & innovation	0.85	-0.74	-1.24	1.13
A respectful, ethical system	-0.34	-0.90	0.52	0.72
Comprehensiveness	-0.05	0.55	0.89	-1.38
Equity	0.09	-0.04	**2.17**	-2.22
People & family centered	-0.45	-1.31	-0.09	**1.85**
Promoting wellness & strengthening prevention	-0.08	-0.38	-0.97	**1.43**
Providing for future generations	0.32	-0.70	-0.88	**1.26**
Public voice & community engagement	0.58	-1.35	0.15	0.62
Quality & safety	**-1.65**	**0.70**	**1.37**	**-0.42**
Recognize social/environ influences shape health	0.72	-1.25	-1.46	**2.00**
Responsible spending	0.07	**1.99**	-0.21	-1.86
Shared responsibility	0.01	0.04	-0.15	0.10
Taking the long-term view	0.49	0.14	-0.77	0.14
Transparency & accountability	-0.62	0.10	1.12	-0.60
Value for money	0.08	**3.14**	-0.45	-2.76
Class size	0.43	0.25	0.17	0.14
Standard deviation	**0.60**	1.20	1.01	1.45

Table 10: Results from 3-Class SALC Cluster Analysis of Derived HB Utilities (Model B2; K=3, S=3)

	Preference parameters		
	3-class SALC cluster model		
Items	Value for the Money	Future	Equity
A culture of reflective improvement & innovation	-1.72	-0.42	-2.17
A respectful, ethical system	-0.27	0.73	0.73
Comprehensiveness	-0.14	-1.29	0.22
Equity	-0.60	-1.61	**1.05**
People & family centered	0.08	**1.91**	0.98
Promoting wellness & strengthening prevention	0.51	**1.66**	0.21
Providing for future generations	0.10	**1.30**	-0.13
Public voice & community engagement	-2.05	-0.68	-1.30
Quality & safety	**1.99**	**1.06**	**2.87**
Recognize social/environ influences shape health	-1.22	**1.04**	-1.52
Responsible spending	**1.57**	-1.01	-0.08
Shared responsibility	-0.77	-0.78	-0.83
Taking the long-term view	0.17	0.15	-0.57
Transparency & accountability	0.05	-0.56	0.72
Value for money	**2.30**	-1.49	-0.19
Class size	0.42	0.33	0.25
Standard deviation	1.25	1.19	1.23

Results from these B-type analyses are similar to those obtained from the corresponding A-type analyses shown earlier. Specifically,

- Results from the 3- and 4-class LC analyses performed directly on the choice responses (A1 analyses) were similar to results from the corresponding LC cluster analyses performed on HB utilities (B1 analyses), in that each of these segmentations included a relatively large "low scale" class.
- Results from the 3-class SALC cluster model (A2) were very similar to those obtained from the 3-class SALC choice model (B2) estimated directly on the choice responses (with S=3 scale classes in both cases). In both cases the resulting segments provide strong support for Flynn's posited segments. This suggests that the SALC model can produce meaningful results when used to cluster on HB utilities.

The one difference in results obtained by the A- and B-type analyses was the following:

- Unlike the A1 results where all segments had *positive* utilities for "Quality & safety" (which overall is the item chosen as Best more than any other item) and negative utilities for "A culture of reflective improvement & innovation" and "Public voice & community engagement," these B1 analyses include segments that had both negative and positive utilities for each of these items.

Evidence that this surprising result is due to a *scale variance confound* comes from an examination of Latent GOLD's "Loadings" output (Table 11) from the 3-class SALC cluster model (Model B2). These loadings relate each of the 15 items to the separate *preference* and *scale* latent variables.[3] In particular, Table 11 shows that unlike the other highlighted items, which load primarily on *preference*, the "Quality & safety" item loads more heavily on *scale*, indicating that most of the variation in this item is due to *scale* heterogeneity. This suggests that when scale is intermingled with preference (as in model B1), the resulting segments are influenced by the relatively large *overall* variance in the "Quality & safety" item, without regard to the fact that its variation is mostly scale heterogeneity.

In contrast, when segments are derived using an SALC model (B2), only the *preference* heterogeneity is used in determining the (preference) segments, in which case the contribution of the "Quality & safety" item to the preference segments is substantially reduced.

Table 11: Loadings Obtained from the 3-Class SALC Cluster Model (Model B2)

	Loadings	
All Items	**Preference**	**Scale**
A culture of reflective improvement & innovation	0.42	0.42
A respectful, ethical system	0.31	0.10
Comprehensiveness	0.38	0.13
Equity	**0.42**	0.11
People & family centered	0.37	0.21
Promoting wellness & strengthening prevention	0.30	0.21
Providing for future generations	0.40	0.15
Public voice & community engagement	0.32	0.39
Quality & safety	0.37	**0.52**
Recognize social/environ influences shape health	**0.58**	0.15
Responsible spending	**0.67**	0.09
Shared responsibility	0.02	0.27
Taking the long-term view	0.21	0.01
Transparency & accountability	0.31	0.01
Value for money	**0.70**	0.09

Comparing Table 10 with Table 8, we see that the SALC model was not only able to eliminate the "low scale" class, but it also removed this scale variance confound associated with the Quality and safety item, resulting in more meaningful segments that confirmed Flynn's hypothesis.

Summary of Best-Worst Responses with and without Scale Adjustment

Comparing Table 10 with Table 5 we see that the segmentation obtained from the SALC Cluster model based on derived HB utilities (B2 model) was quite similar to the segmentation obtained from the best practice SALC choice model directly from the Best-

[3] For further details on these loadings, which are analogous to *factor loadings* in factor analysis, see Magidson and Vermunt (2003), and Vermunt and Magidson (2004).

Worst choices (A2 model). Not only do both segmentations support the Flynn hypothesis, but the two segmentations are in agreement with respect to the sizes of the 3 segments.

Table 12: Summary of SALC Modeling Impact

	Best-Worst Choice Responses	Derived HB Utilities	
	Remove low scale confound?	Remove low scale confound?	Remove Variance confound?
Without scale adjustment (LC)	No	No	No
With scale adjustment (SALC)	Yes	Yes	Yes

Moreover, Tables 13 and 14 below show that the segmentations obtained from the A2 and B2 models assign respondents to the same segment and the same scale class 88% and 87% of the time respectively, which is about what would be expected if the segmentations were subject to 10–15% misclassification due to chance.

Table 13: Crosstabulation of segment assignments based on SALC models obtained by analyzing HB Utilities (rows) vs. Best-Worst responses (columns). Overall, 88% of respondents were assigned to same segment.

Preference Segment Crosstabulation					
SALC segments based on HB Utilities		SALC segments (Preference Classes) based on Best-Worst Responses			
		1	2	3	Total
Segments (Preference Classes)	1	75	1	2	78
	2	10	64	2	76
	3	5	5	40	50
Total		90	70	44	204

Table 14: Crosstabulation of scale class assignments based on SALC models obtained by analyzing HB Utilities (rows) vs. Best-Worst responses (columns). Overall, 87% of respondents were assigned to the same scale class.

Scale Class Crosstabulation					
Scale classes from SALC analysis of HB Utilities		Scale classes from SALC analysis of Best-Worst Responses			
		High	Middle	Low	Total
Scale classes	High	33	9	0	42
	Middle	7	84	2	93
	Low	0	9	60	69
Total		40	102	62	204

Thus, we conclude that the SALC model can produce meaningful segments not only when based on Best-Worst choices (the best practice/gold standard approach A2), but also when used to cluster on HB utilities derived from the Best-Worst choices (B2). As mentioned earlier, while approach A2 is preferred for theoretical reasons, when the basis variables contain both choice and non-choice data, the input data file is easier to set up with the B2 approach. In the next section we compare the performance of the A2 and B2 approaches in an application where the basis variables consist of both choice (Best-Worst) and non-choice (attitudinal) data.

Example 2: Analysis of Choice and Non-Choice Data

Practitioners often prefer to include both choice as well as non-choice variables in their segmentation. As mentioned above, best practice for conducting such an analysis is to construct a rather complex data file consisting of responses to both the choice and non-choice questions (see Appendix C) and to use an A2-type model that analyzes both the choice and non-choice responses simultaneously. The problem with this approach is that the complexity of setting up the necessary data file makes such analysis somewhat difficult to conduct.

On the other hand, the results obtained by segmenting the Best-Worst data from the Australian Health Care Reform study suggested that the use of the SALC model B2 to cluster on HB utilities yields meaningful segments quite similar to those obtained using the best practice A2 analysis where SALC is applied directly to the Best-Worst choice responses. In this section we investigate whether meaningful segments might also be attainable from the simpler data setup, when the basis variables consist of not only HB utilities derived from Best-Worst choice data, but also non-choice data such as attitudinal variables.

The data for our analyses consists of one Best-Worst task and one attitudinal battery from the Global Travel Retail Industry Cross-Category Segmentation Study conducted in 2012 by M1ndSet. This study focused on the shopping behaviors and attitudes of N = 3,433 international travelers shopping at airport duty-free shops. For more details of this study see Appendix B.

Our analysis proceeds as follows:

1. Validation of earlier results. We begin by analyzing only the Best-Worst data to attempt to validate the results obtained from our earlier LC analyses (A1 and B1 analyses). Namely,
    - Do the A1 and B1 approaches provide evidence for a "low scale" class or was our earlier result specific to that earlier data set?
    - Do the A2 and B2 SALC analyses provide more meaningful segments than A1 and B1, respectively?
2. Extension of SALC models to include both choice and non-choice data.

Validation of Earlier Segmentation Results Based on Best-Worst Data Only

Results from both the A1 and B1 LC analyses of the Best-Worst data from the Global Travel study again supported the emergence of a "low scale" class. For the A1 analyses (conducted on only the Best-Worst responses), a "low scale" class emerged in 3-class, 4-

class, and 5-class models, representing 36%, 27%, and 24% of the sample, respectively. For the corresponding B1 analyses, using LC to cluster the HB Best-Worst utilities, again “low scale” classes emerged, representing 50%, 63%, and 33% of the population respectively. However, there was a substantial amount of dissimilarity between the two 5-class (A1 vs. B1) solutions; overall, only 47% of respondents were classified into the same class. Correspondence was similarly poor for the 3- and 4-class solutions.

Repeating the 5-class analyses using the SALC models (A2 and B2 analyses with K=5 preference classes and S=2 scale classes), the parameters for the resulting preference classes again had similar standard deviations, consistent with the removal of the scale confound. In contrast to the dissimilarity revealed by cross-tabulating the segment assignments from the A1 and B1 analyses, the 5 *preference* classes showed a high degree of similarity, with a much higher percentage of respondents (67%) assigned to the same class. These results validate our earlier conclusion that classes obtained from clustering on the HB utilities using the SALC model provide more similar segments than clustering these utilities with the standard LC model.

Extension to the Choice and Non-Choice Data

We will now compare segmentations resulting from the A2 and B2 SALC models where the basis variables consist of both choice (Best-Worst) and non-choice (attitudinal) data. In particular, we focus on the question of whether the B2 SALC approach can yield a segmentation comparable to the best practice but more complex data setup associated with the A2 SALC approach. For simplicity, we settled on comparison of SALC segmentations with K=5 preference classes and S=2 scale classes.

Since the scale adjustment mechanism is likely different for the choice and non-choice variables, separate scale-adjustments can be applied to each set of variables as in Magidson et al. (2009).[4] For simplicity, we decided to apply the SALC model only to the choice variables, but we also experimented with separate variable weightings for the choice and non-choice variables. See Appendix D for the Latent GOLD syntax to accomplish this for the B2 SALC model.

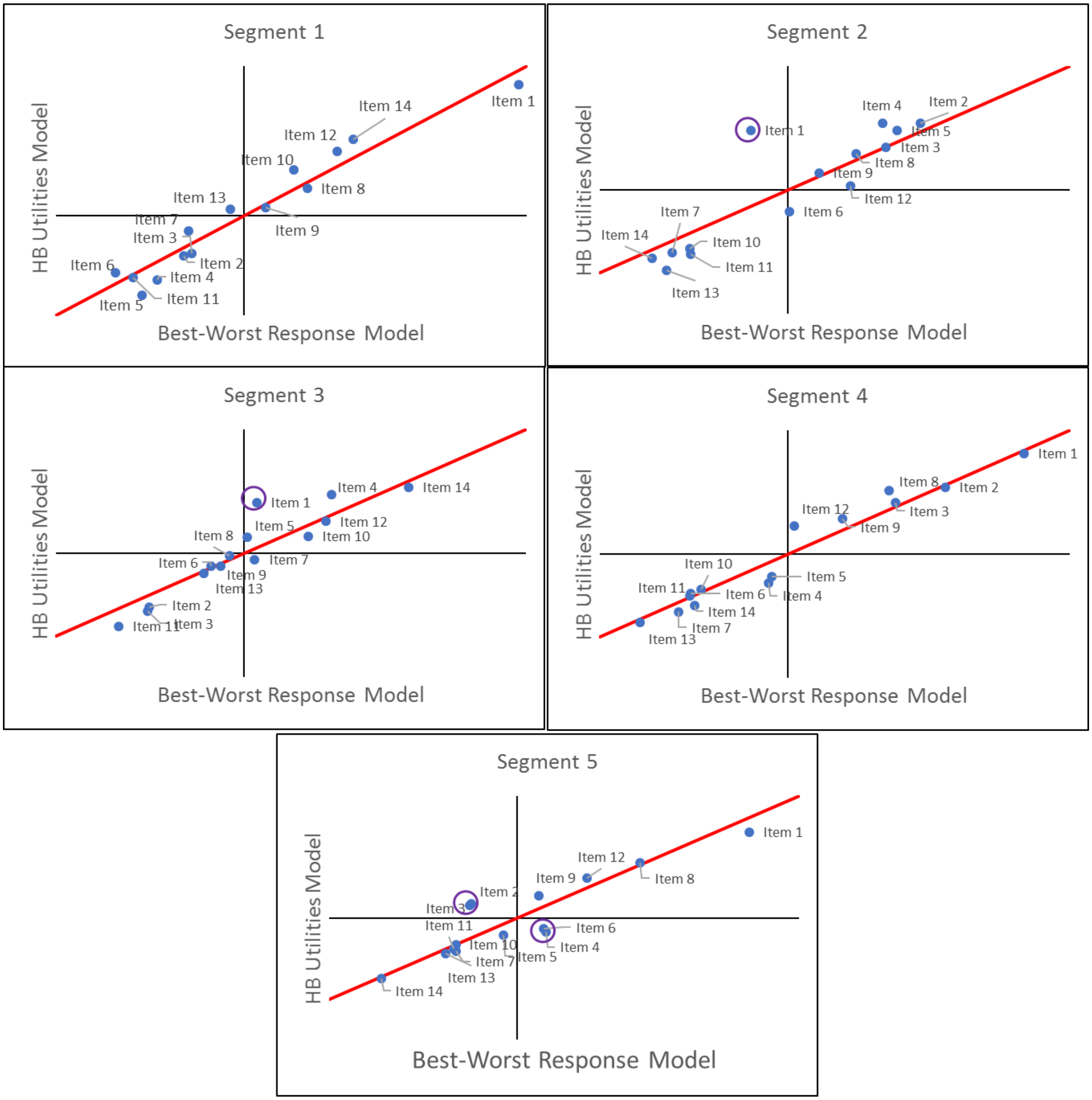
Class assignments obtained from the A2 and B2 SALC models show a high degree of association. The scale class to which the respondents are assigned by these models agrees over 80% of the time. Moreover, assignment to each of the five preference segments is in agreement (same modal class), with the overall rate of agreement (68%) being much higher than the corresponding solution without the use of scale classes (47% agreement between A1 and B1). For more details see Appendix B (Tables B4 and B5).

More interesting to practitioners, the segment profiles also show high agreement between the A2 and B2 SALC approaches on the choice and non-choice data. Figure 1 shows the mean SALC A2 choice model parameters (horizontal axes) plotted against the corresponding mean SALC B2 HB parameters (vertical axes) separately by segment. The red line is the 1:1 diagonal reference line. If the segments had the same exact parameters, the points in the plots would fall along this line. For the most part there is a strong association between the parameters of both segmentation methods. The average R-square is .87. There are some anomalies: the circled points in segment plots 2, 3, and 5 are Best-

[4] See pages 101–102 of that paper for the specification of that “fused” model using an earlier version of Latent GOLD.

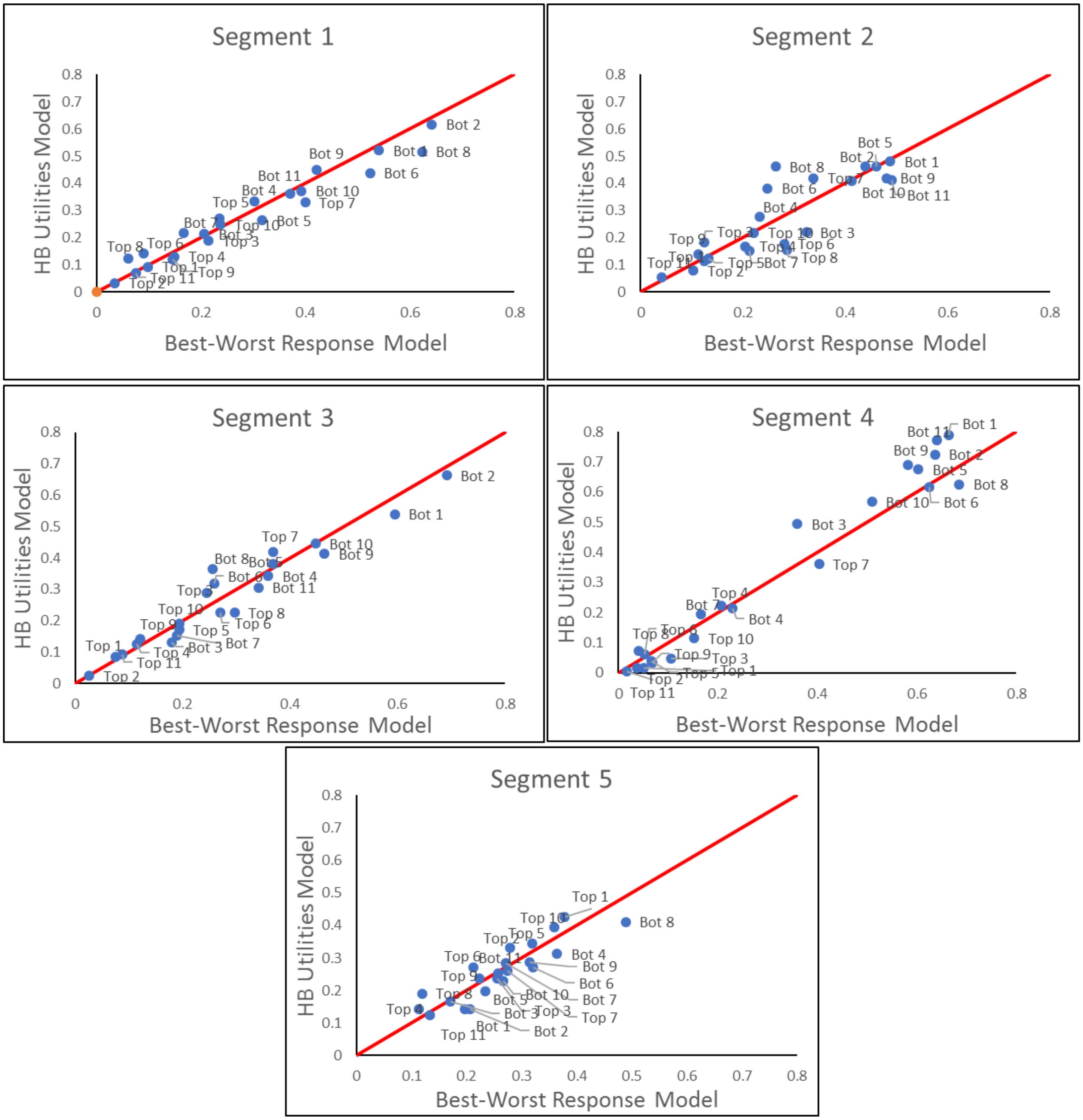
Worst items where the signs are different between the two methods and the distance of the point from the diagonal red line seems large. The circled points account for 6 of the 70 points across all 5 segments (14 items * 5 segments).

Figure 1: Cross Method Best-Worst Parameters by Segment



The scatter plots of the *non-choice* attitudinal statements (Figure 2) show a similar pattern of association. Figure 2 plots the top 2 and bottom 2 box proportions for the 11 attitudinal statements (hence, each plot has 22 points labeled “Top XX” and “Bot XX,” where XX is the attitudinal statement number (from Table C2 in Appendix C). The horizontal axes represent the segment specific profiles from the SALC choice segmentation, and the vertical axes show the SALC clustering of derived HB parameters. The average R-square of these plots is .86. Segments 1 and 3 show a strong association across methods. Segment 4 appears to have a strong association, except for the bottom 2 box proportions in the upper right-hand corner of the plot. Segment 2's profile points seem to show more spread away from the diagonal than segments 1 and 3, indicating there is more of a difference in this segment profile. Segment 5 shows a relatively strong association, but the fact that the points are all clustered in the range of 0.2 to 0.5 suggest poor differentiation on the non-choice variables in this segment. This may be a result of the last segment being a "catch-all," or "outlier," category as the BIC statistic provides evidence for more than 5 segments.

Figure 2: Cross Method Non-Choice Attitudinal Statements by Segment



Weighting the Segmentation Variables

Overall, we have found evidence that meaningful segments can be obtained from the application of SALC models to HB-derived choice parameters, even when combined with non-choice variables. However, as our investigation progressed with the simultaneous analysis of the choice and non-choice variables, we noticed that the segments, from both the A2 and B2 approaches, appeared to be influenced more heavily by the *choice* as opposed to the *non-choice* variables. This led to the development of the second advance in Latent GOLD 6.0—a variable weighting capability within the latent class framework that enables the researcher to differentially weight sets of variables during segment extraction. We examined results obtained by weighting the choice variables less heavily than the non-choice variables.

To illustrate the impact of differential variable weighting, we estimated a simple 2-class LC cluster model based on the HB utilities data combined with the non-choice data. For purposes of comparison, we estimated this model with and without variable weights. The unweighted analysis was conducted using the same weights, 1.0, for all choice and non-choice variables, as indicated in the left side of Table 15. In the right side of Table 15, results of the LC cluster analysis are presented where we down-weight the 14 Best-Worst utilities, using a weight of 0.5. See Appendix D for the Latent GOLD syntax for these models.

Table 15: Impact of differential weighting of the choice and attitudinal variables based on a 2-class SALC cluster (B2) segmentation of the derived HB parameters.

All Variables Weighted 1.0						HB Utilities Down-weighted to 0.5					
Choice Utilities	Weight = 1.0		Attitudinal Variables	Weight = 1.0		Choice Utilities	Weight = 0.50		Attitudinal Variables	Weight = 1.0	
	p-value	R^2		p-value	R^2		p-value	R^2		p-value	R^2
asc_1	0.79	0.10	nq19_1	5.2E-05	0.01	asc_1	0.048	0.00	nq19_1	2.7E-73	0.35
asc_2	3.5e-388	0.54	nq19_2	0.61	0.00	asc_2	2.4E-09	0.03	nq19_2	3.4E-68	0.23
asc_3	2.7e-364	0.52	nq19_3	7.4E-21	0.04	asc_3	2.6E-07	0.02	nq19_3	2.1E-51	0.13
asc_4	0.00024	0.01	nq19_4	1.3E-10	0.02	asc_4	0.04	0.00	nq19_4	2.0E-09	0.01
asc_5	2.9E-53	0.09	nq19_5	1.8E-26	0.05	asc_5	0.67	0.00	nq19_5	6.9E-61	0.18
asc_6	0.48	0.02	nq19_6	5.8E-09	0.01	asc_6	0.0002	0.01	nq19_6	3.9E-39	0.09
asc_7	6.9E-66	0.13	nq19_7	0.064	0.00	asc_7	1.6E-06	0.02	nq19_7	1.9E-05	0.01
asc_8	1.5E-61	0.15	nq19_8	1.3E-07	0.01	asc_8	0.00086	0.01	nq19_8	2.0E-16	0.03
asc_9	3.6E-39	0.07	nq19_9	3.0E-07	0.01	asc_9	3.8E-05	0.01	nq19_9	2.9E-57	0.14
asc_10	2.5E-225	0.33	nq19_10	1.4E-05	0.01	asc_10	0.072	0.00	nq19_10	4.3E-46	0.10
asc_11	5.4E-04	0.13	nq19_11	2.9E-28	0.05	asc_11	0.18	0.00	nq19_11	6.6E-68	0.19
asc_12	5.1E-37	0.09				asc_12	0.76	0.00			
asc_13	4.1E-150	0.26				asc_13	2.8E-08	0.02			
asc_14	1.7E-194	0.29				asc_14	0.60	0.00			
Avg. R^2		0.20	Avg. R^2		0.02	Avg. R^2		0.01	Avg. R^2		0.13

In particular, Table 15 shows the impact on the p-value and R-square statistics of down-weighting the HB parameters. In the left-most set of columns, the pink/shaded cells under the column labeled “HB variables” are the Best-Worst items that had a highly significant p-value (< 1E-50) and R-square value > 0.10 when the weights for both the HB utilities and non-choice attitudinal variables were set to 1.0 (The R-square column indicates the degree to which this 2-segment solution predicts the original value of the HB utilities). In the HB variables' p-values column, many of the values are quite significant as indicated by their very small magnitude. The left-most set of "Attitudinal Variables" columns indicate the p-value and R-square for the attitudinal variables. Notice there are still some significant p-values, but the R-square values are all less than 0.10.

The right-most set of columns show the results from the 2-segment LC cluster solution where we down-weighted the HB utilities with a weight of 0.50. Notice under the "HB Variables" p-value column that most of the "ASC" parameters have reduced significance (higher p-values) and all of their R-square values are less than 0.10. This is a result of the down-weighting. Examination of the right-most set of "Attitude Variables" columns show some yellow/shaded cells. These cells are the attitudinal statements that are now highly significant (very low p-values) and with R-square values greater than 0.10. All the attitudinal statements have become more significant than before.

Notice that by down-weighting the HB utilities by .50, we have illustrated both extremes—from one extreme where the HB variables dominate the segments to the other extreme where the attitudinal variables dominate. Down-weighting the HB utilities by a value between 0.50 and 1.0 will produce a more balanced mix between the significant HB utilities and attitudinal variables. While these results demonstrate the impact of weighting on a 2-segment solution, what is more interesting is to examine the impact graphically in our 2-scale class, 5 preference segment solution.

It is our observation that, in the Duty Free A2 SALC Best-Worst response model (K=5 preference segments, S=2 scale classes), the choice variables impact the solution more than the non-choice variables. Therefore, rather than changing the A2 SALC Best-Worst response model, we focus our attention instead on the B2 SALC HB utilities model. And rather than weighting up the HB utilities in the B2 SALC HB utilities model, we down-weight the non-choice variables to make the B2 SALC HB utilities model more comparable to the A2 SALC Best-Worst response model. We then compare the results of the down-weighted B2 SALC HB utilities model to the unweighted A2 SALC Best-Worst response model.

In the plots below (Figure 3) we plot the B2 SALC HB utilities model's attitudinal variables weighted down by 0.5 against the unweighted A2 SALC Best-Worst response model (5 preference segments, 2 scale classes). The horizontal axes are the original unweighted A2 SALC Best-Worst response model solution's attitudinal variable profiles (e.g., the Top XX and Bot XX profile proportions). The vertical axes are for the same set of variables down-weighted by 0.5 from the B2 SALC HB utilities model. Except for segment 5, we see the trendline of the scatter points moving in a slight clockwise direction relative to the diagonal. This is most clearly seen in the segment 4 profile shown in Figure 3.

To examine the change in results in the B2 SALC utilities model when we down-weight the non-choice attitudinal variables, we next plot the unweighted B2 SALC utilities model non-choice attitudinal variable profiles against the down-weighted B2 SALC utilities model (weight = 0.5). Figures 4 and 5 depict the change in non-choice attitudinal variable impacts as a result of down-weighting. The horizontal axes are the unweighted B2 SALC utilities model non-choice attitudinal variable top and bottom 2 box proportions. The vertical axes are the down-weighted B2 SALC utilities model non-choice attitudinal variable proportions.

Here one can see more clearly the impact of down-weighting by examining the extremes of the plot for segment 4 (Figure 4). The lower left-hand corner collection of points (proportions of Top XX and Bot XX variables) is above the diagonal while the upper right-hand corner collection of points is below the diagonal. This shows a reduction in the differentiation (variance) among the attitudinal variables as a result of down-weighting the non-choice variables.

Segment 5 (see Figure 5) shows something unexpected. It appears the profile of non-choice variables has become more random than before weighting. This was also seen in the plot of segment 5 in Figure 3. While unexpected, the pattern may be the result of segment 5 being a segment that is catching all the "outlier" heterogeneity that exists within the solution because of our arbitrary selection of a 5-segment solution (i.e., when the optimal number of segments, based on the BIC, is greater than 20).

In summary, differential variable weighting, as a new tool in the practitioner's analysis kit, appears to be useful. However, the fact that the amount of down-weighting must be determined by the practitioner means that it is not an off-the-shelf solution, but a "tuning" parameter that will take some time to get used to. Here we examined only a couple of potential uses for variable weighting. Seeing such a changed pattern in segment 5 as compared to the remaining segments leads us to suggest that the impact of weighting variables needs to be explored further.

Figure 3: Cross Method Non-Choice Attitudinal Statements by Segment Down-Weighted to 0.50

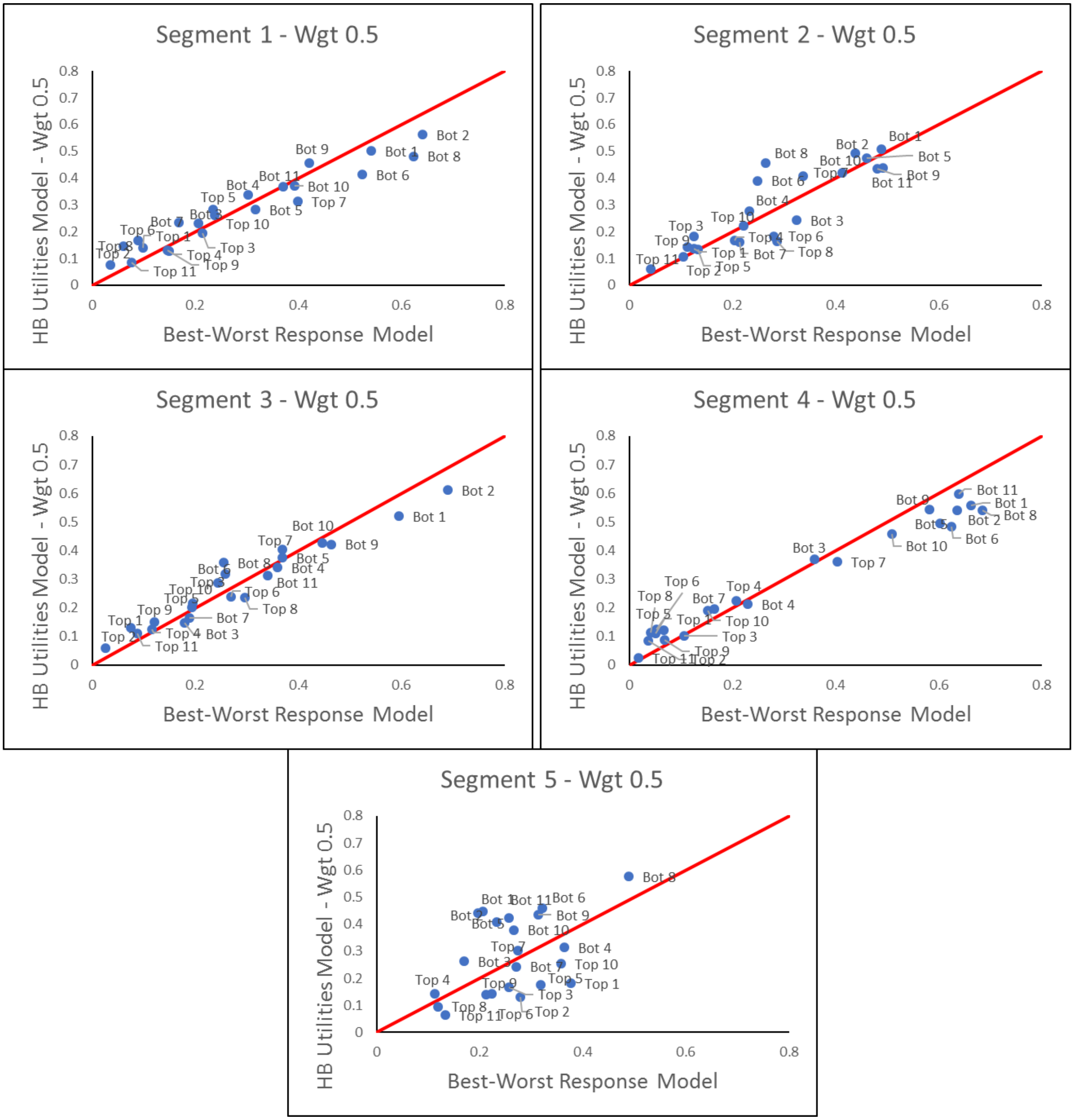


Figure 4: Segment 4 Comparison of Weighted to Non-Weighted Attitudinal Variables

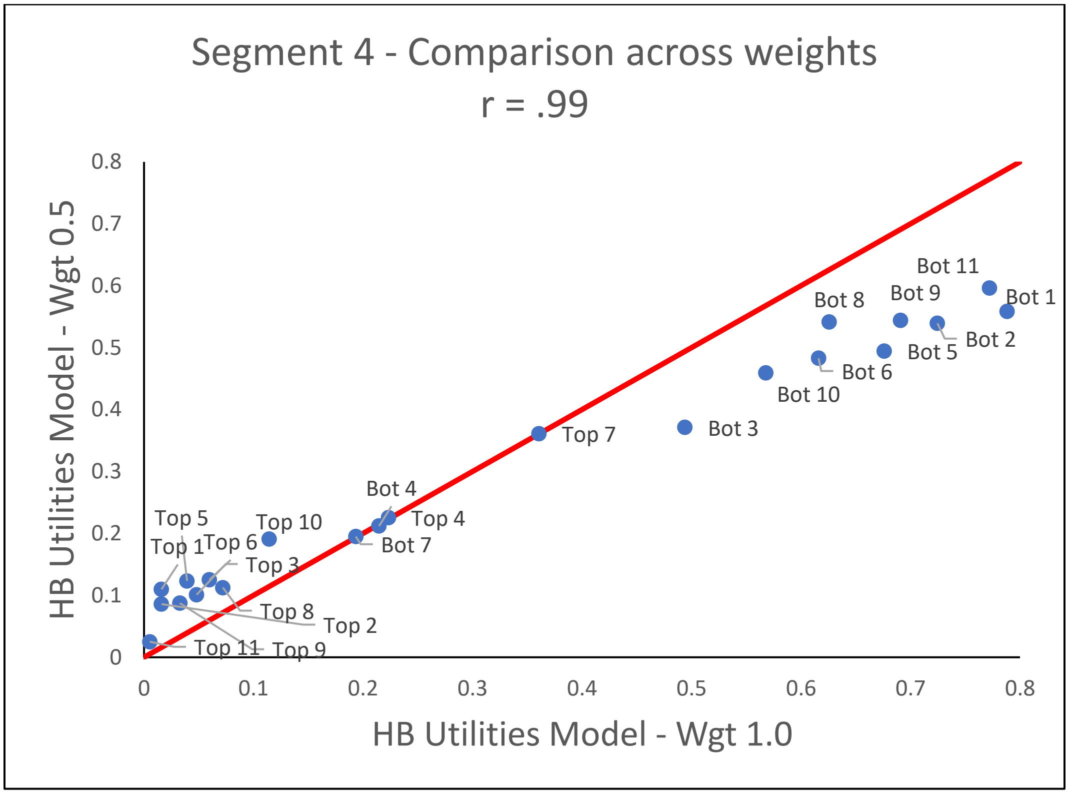
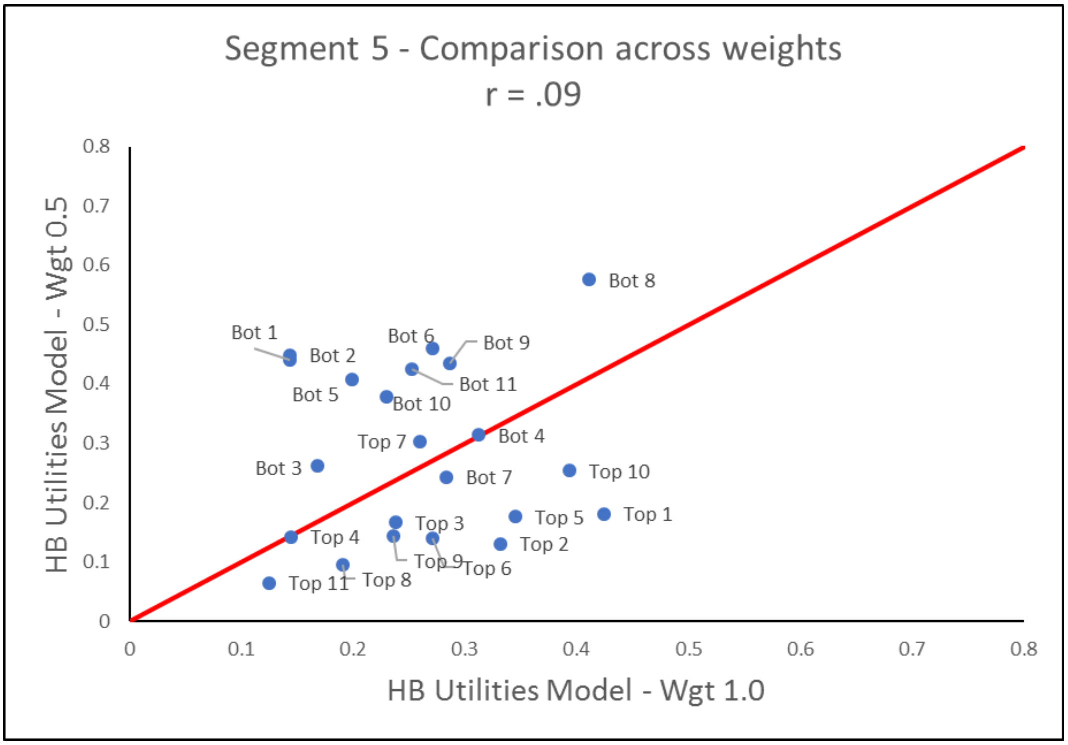


Figure 5: Segment 5 Comparison of Weighted to Non-Weighted Attitudinal Variables



SUMMARY AND CONCLUSIONS

We reconsidered the usefulness of HB utilities (or lack thereof) as basis variables with latent class segmentation. We undertook this task not because we are die-hard HB fans, but because of the possibility of obtaining an effective segmentation, conveniently, with data already at hand. Despite the bad press and many strong cautions against using HB utilities in latent class segmentations due to the frequent occurrence of "surprising results," many practitioners continue to use HB utilities with LC segmentation anyway because they already have these at hand. Moreover, it would be very straightforward to combine HB utilities with an attitudinal battery of items in a simple rectangular (cases by variables) file to perform simultaneous segmentation of choice and non-choice data much more easily than using the best practice approach.

Thus, we explored whether such surprising results could be the result of scale confounds, and could possibly be eliminated using scale-adjusted latent class (SALC) models to perform the segmentation. For our evaluation we utilized two Best-Worst datasets, the second being supplemented with attitudinal data. The first application in this paper used the Best-Worst data from the Australian Health Reform Study in a confirmatory application to confirm the existence of three posited segments. We compared the standard 3-class LC models with 3-class SALC models, where both models were estimated in 2 different ways:

1. "A-type" analyses—using LC and SALC *choice* models to segment Best-Worst responses directly, and
2. "B-type" analyses—using LC and SALC *cluster* models to obtain segments based on individual-level HB parameters, derived from the Best-Worst responses and treated as continuous variables.

Both models A1 and B1 resulted in segments that were confounded with scale, and an additional *variance* confound problem was encountered in the LC segmentation (B1). However, the preference classes obtained by SALC (models A2 and B2) were no longer confounded; both provided strong confirmation for Flynn's three posited segments. The problem evidenced in the B1 analysis was eliminated along with the scale confound by the SALC model in the B2 analysis.

Prior to moving on to our simultaneous utility and attitudinal segmentation with our second application, we repeated the A1 and B1 analyses with these data using just the HB utilities, and reproduced the results obtained in the A2 and B2 analyses, again successfully eliminating the surprising result that occurred during the B1 stage of analysis. We thus conclude that the SALC model when applied to HB utilities can yield a more effective segmentation.

We then proceeded with our simultaneous segmentation, and achieved a successful result using the SALC model supplemented by the ability to down-weight one group of variables (choice variables to which scale-adjustment was also applied) relative to another group of variables (attitudinal variables to which scale-adjustment was not applied) as part of the modeling, to allow attitudinal basis variables to play a greater role in the segmentation.[5]

[5] As an example of a Latent GOLD 6.0 syntax where the SALC model is used with and without differential weighting, see Appendix D.

Down-weighting (or more generally, differential weighting), along with the possible use of scale factors with either one or both sets of variables (recall Table 15), was found to be an interesting new tool for the practitioner's toolkit, However, the subjective nature of this weighting tool provides a challenge for its future use.

Overall, our research reinforces the many warnings against the current widespread use of clustering on HB utilities, due to the resulting scale confounds that are likely to yield misleading interpretations from the segments thus obtained. However, returning to our original question, "Would the inclusion of scale classes in a LC model enable practitioners to obtain meaningful segments based on HB derived parameters?," our results here suggest that the use of the Scale-Adjusted LC (SALC) model to cluster on zero-centered HB utilities,[6] with or without the inclusion of attitudinal or other additional basis variables, can yield meaningful segments similar to those obtained from the best-practice SALC models applied directly to individual responses.

Thomas C. Eagle

Jay Magidson

Appendix A. The Scale-Adjusted Latent Class (SALC) Model in Latent GOLD

The SALC model was proposed by Magidson and Vermunt (2007)[7] for categorical (including choice) response variables and estimated using the syntax module of Latent GOLD® version 4.5. In the SALC model, a single discrete latent variable is assumed to be the source of the preference parameters, its categories called "preference classes," and a separate latent variable is used to model the scale parameters.

This SALC model was extended in the LG version 5.0 syntax (Vermunt and Magidson, 2013) by embedding it in a very general log-linear sub-model that allows scale factors to be modeled as a function of both observed and latent variables, and allows observed covariates to be included as predictors of these latent variables. Let the parameter $\beta_{j.ks}$ denote the utility for attribute j, for respondents in preference class *k* and scale class *s*. SALC uses a log-*bilinear* form to specify the random utility model, $\beta_{j.ks}$ being decomposed into the product of separate *preference* and *scale* components:

$$\beta_{j.ks} = \exp(\lambda_s - \lambda_0)\beta_{j.k1}$$

where a log-scale factor λ_s is estimated simultaneously with the preference parameters $\beta_{j.k1}$.

[6] It is our belief that the *only* way that clustering of HB utilities can achieve consistency with the gold standard is to perform SALC clustering on zero-centered HB utilities, as we did here. We are conducting additional research to test this belief.

[7] Development of the SALC model was motivated by Louviere and Eagle (2006).

Use of a log-linear structure to estimate the log-scale parameters guarantees that estimates for the scale factors are always positive. For purposes of identification, λ_s is determined relative to a fixed reference point λ_0, and can be modeled using either a group-level (discrete)[8] or an individual-level (continuous) latent variable (for further details see Vermunt and Magidson, 2013). In contrast, the standard LC choice model does not separate scale from preference, and thus cannot be distinguished from a model that assumes no scale heterogeneity (i.e., $\lambda_s = \lambda_0$ for all s).

SALC was further extended in LG version 6.0 (forthcoming 2020) for use in LC cluster models with continuous indicators such as HB individual-level parameters, as well as in LC regression models such as Poisson, negative binomial, and zero-inflated models.

To show that the SALC model yields a better fit to the data, Table A1 below displays the log-likelihood (LL), Bayesian Information Criteria (BIC), and number of parameters (Npar) for the standard LC choice models with K = 1 to 4 classes, and for some SALC choice models. In order to explain more heterogeneity in data, the standard exploratory LC modeling approach is to increase the number of classes. For example, Table A1 shows that the increase from 3 to 4 classes improves the log-likelihood by 192 (from LL = -10585 for the 3-class model to -10393 for the 4-class model). The cost in terms of model complexity for this improvement is the addition of 15 parameters (from Npar =44 in the 3-class model to Npar = 59 in the 4-class model).

Alternatively, the 3-class SALC model adds only 4 additional parameters[9] to the 44 parameters in the 3-class model, but fits the data better[10] than the 4-class LC model, which adds 15 parameters! Overall, the 3-class SALC model fits best (lowest BIC) among the models listed in Table A1.

Table A1: Model fit Comparison where log-scale factors are modeled as a function of a *discrete* latent variable (the categories of which are called "scale classes").

Standard Latent Class (LC) Models (no adjustment for scale)				SALC Models (with 3 scale classes)			
Model	LL	BIC	Npar	Model	LL	BIC	Npar
1-class	-11246	22566	14	1-class SALC	-11035	22166	18
2-class	-10815	21784	29	2-class SALC	-10573	21321	33
3-class	-10585	21405	44	3-class SALC	-10360	20974	48
4-class	-10393	21100	59				

Table A2 compares various 3-class SALC models, including a 3-class SALC model that uses a continuous latent variable to model the log-scale factors. We see that the best fit (lowest BIC) occurs with the discrete form of the SALC model with 3 scale classes (BIC =

[8] For identification in the case of a discrete latent variable with S scale classes, by default Latent GOLD uses the first scale class (s=1) as the reference class, by setting $\lambda_1 = \lambda_0 = 0$.

[9] These 4 additional parameters consist of 2 *size* parameters for the first two scale classes (since respondents are classified into one of the 3 scale classes with probability 1, the probability of respondents being in the 3rd scale class is determined as 1 minus the probability of being in scale classes 1 or 2), plus the two (relative) scale factors, one associated with the medium scale class (relative to the large scale class), and the other associated with the "low scale" class (relative to the large scale class).

[10] Table A1 shows that LL= -10360 for the 3-class SALC model, which is larger than LL= -10393 for the 4-class LC model.

20974). Note that this model fits better than the SALC model with continuous scale (BIC = 20992). In contrast to the continuous scale, which assumes that scale follows a log-normal distribution, use of the discrete form of the SALC model requires no assumptions about the distribution of scale or the size of the scale classes being estimated from the data simultaneously with the other model parameters[11].

Table A2: Model Fit Comparison of Various Discrete and Continuous Latent Scale Variables

3-Class SALC	LL	BIC(LL)	Npar
1 scale class	-10585	21405	44
2 scale classes	-10394	21033	46
3 scale classes	-10360	20974	48
4+ scale classes	-10360	N/A *	N/A *
Continuous scale	-10376	20992	45

Note also that the LL does not change when the number of scale classes is increased beyond 3 (LL = -10360 for both the 3-class and 4-class SALC models). This means that the 3-class SALC model reaches a saturation point with 3 scale classes. We have observed this saturation phenomenon in SALC models with other data as well.[12] The SALC structure is very restrictive and thus does not pick up much heterogeneity, except for proportionality of all parameters simultaneously, which occurs only when pure scale heterogeneity is present.

APPENDIX B: DESCRIPTION AND SOME RESULTS OF THE CHOICE AND NON-CHOICE DUTY-FREE DATA

The data for this portion of the analyses is from the Global Travel Retail Industry Cross-Category Segmentation Study conducted in 2012 by M1ndSet. This study focused on the shopping behaviors and attitudes of international travelers shopping, or potentially shopping, at airport duty-free shops. International air travelers were recruited at 28 worldwide airports. The recruits were directed to complete an online survey about their reasons for, preferences for, and attitudes towards shopping at duty-free shops as well as shopping at more traditional shopping centers. The sample consisted of 4,519 respondents who completed two different Best-Worst tasks in addition to completing survey questions regarding attitudes, behaviors, and socio-demographic characteristics. Respondent groups included frequent and infrequent flyers, business versus leisure flyers, and covered the major regions of the world: Europe, Asia/Pacific, Middle East, North and South America.

Our focus here is on one of the Best-Worst tasks and one attitudinal battery. The Best-Worst task included 14 items. It was comprised of a balanced incomplete block design

[11] By not making any distributional assumptions, the discrete form of the SALC model differs from the G-MNL model which attempts to achieve identification by assuming different distributions for the scale parameter and preference parameters. As a result, unlike the SALC model, G-MNL yields weak identification (see Hess and Rose, 2012). Because no distribution is assumed, SALC explains all associations between the utilities, which includes linear associations (correlations), as well as non-linear associations that may exist in the data. A standard mixed/HB logit or LC model can't separate preference from scale because there is no separate scale term in the model. Note also that HB and similar approaches make assumptions that imply the associations between utilities are linear, but this need not be true.

[12] Saturation of this type and conditions for achieving it has been discussed by Lindsay et al. (1991), where semiparametric estimation of the Rasch model using latent classes achieves exactly the same model fit (and same number of degrees of freedom) as the conditional Rasch model.

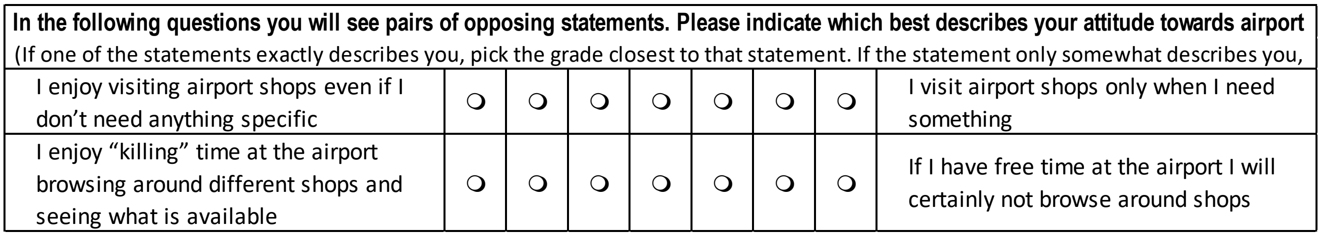
(BIBD) design of 14 tasks with 3 items in each task. Each respondent was asked to select the benefit regarding shopping that was most and least important to them. The full list is in Table B1 below.

Table B1: Best-Worst items

1	Same product as downtown but at cheaper prices
2	Limited/ special editions only found at airports
3	Exclusive products/brands
4	Local products/specialties
5	Well known international product but with a local touch
6	Finding products suitable as gifts/in a gifting packaging
7	Products for immediate consumption/to use during my trip
8	Guaranteed good quality compared to downtown
9	Opportunity to compare/try out different brands at one location
10	Having more time to shop
11	Better advice from sales staff compared to downtown
12	Pleasant shopping environment (nice/clean/big)
13	Shopping alone/on your own
14	To kill time/entertain me before my flight

The attitudinal battery was a series of opposing attitudinal statements regarding the respondents' attitudes towards airport shopping. Figure B1 below shows a subset of the opposing statements and how the respondent were asked to respond to the statements. Table B2 provides the complete list of opposing airport shopping attitudinal statements.

Figure B1: Example of the Opposing Attitudinal Shopping Behavior Attitudes



In the following questions you will see pairs of opposing statements. Please indicate which best describes your attitude towards airport

(If one of the statements exactly describes you, pick the grade closest to that statement. If the statement only somewhat describes you,

I enjoy visiting airport shops even if I don't need anything specific	○	○	○	○	○	○	○	I visit airport shops only when I need something
I enjoy "killing" time at the airport browsing around different shops and seeing what is available	○	○	○	○	○	○	○	If I have free time at the airport I will certainly not browse around shops

Table B2: Complete List of Opposing Shopping Attitudinal Statements

1	I enjoy visiting airport shops even if I don't need anything specific	I visit airport shops only when I need something
2	I enjoy "killing" time at the airport browsing around different shops and seeing what is available	If I have free time at the airport I will certainly not browse around shops
3	Airport shops are among my favourite places to shop when I travel	I much prefer to shop in other places than airport shops during my trip
4	I go to airport shops mainly for convenience reasons (no time to shop elsewhere, easier to find what I need, etc)	I go to airport shops because they offer a great shopping experience
5	I like airport shops that have novelties, promotions, let you try new things, etc.	I'm mostly interested in buying my usual product(s) at airport shops
6	I like to browse around airport shops to try and find bargains, good prices, quantity discounts, etc.	I go to airport shops for other reasons than price (choice, quality, convenience, service, etc.)
7	I use airport shops to avoid having to think about shopping during other moments of my trip	When I travel, I like to visit all kind of shops either at the airport or downtown
8	I usually compare/ know the prices of the products I need and buy at airport shops only if it is cheaper than in downtown shops	I don't look in details at the price differences between downtown and airport shops
9	In general I like to go early at the airports and then have time to shop, read, relax, etc. before my plane leaves.	I don't like to waste time at the airports and try to arrive last minute.
10	If I have foreign money left from my trip, I'll try to spend it at airport shops	If I don't really need something, I prefer to keep/ change back foreign money
11	I'm interested in the selection of products one can find in airport shops	Airport shops do not have the type of products I usually buy.

Some Duty-Free Shopping Results

These data are very heterogeneous. This is evident in the BIC statistics derived from both segmentations. Table B3 shows that the BIC statistic continually improved until we reach between 20 and 25 segments using the derived HB parameters; for the choice data, it continually improved to 30 segments, where we stopped the analysis.

Table B3: BIC Statistics for an Increasing Number of Segments

		HB Utilities Segments	**Best-Worst Response Segments**
		BIC	BIC
No scale classes	5 segments	252,277.4	288,566.3
	10 segments	249,932.2	283,938.8
	20 segments	**248,703.5**	280,444.5
	25 segments	248,743.2	279,823.4
	30 segments	249,023.7	**279,493.1**
2 scale classes	5 preference segments	251,611.9	287,239.2
	10 preference segments	249,607.1	283,093.8
	20 preference segments	**248,539.9**	280,088.1
	25 preference segments	248,570.5	279,533.9
	30 preference segments	248,874.0	**279,156.5**

These results suggest heterogeneity in the data—many segments would be required to optimally segment these data. However, this many segments are not practical for managerial purposes. Typically, practitioners and research managers prefer between 4 to 8 segments. For the sake of displaying results in small tables we arbitrarily decided to present results for the 2-scale class, 5-preference segment solutions. Since there are more preference segments than earlier, and thus more ways for respondents to be misclassified, and because our selection of classes was arbitrary, we might not expect to see the same level of agreement between the solutions as we saw with the Flynn data presented in the first example.

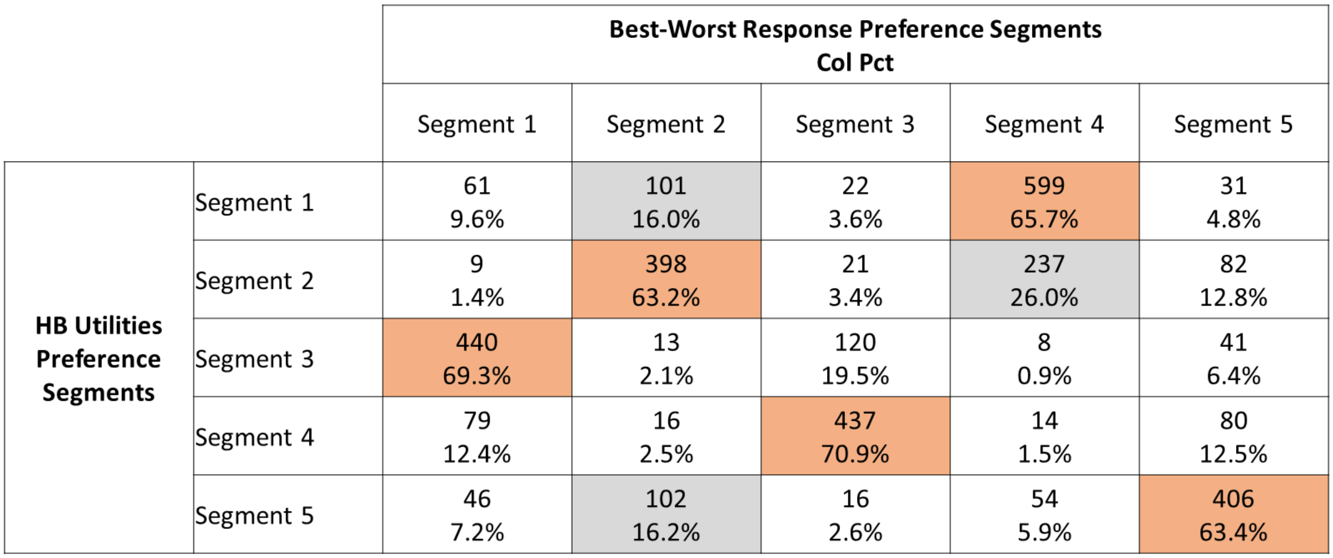
Table B4 gives the cross tabulation of scale class membership for the two solutions. The rows are the 2-scale classes for the latent class clustering of the derived HB parameters. The columns are the 2-scale classes for the latent class choice segmentation. The scale classes identified by Latent Gold are in the row and column headers. The percentages are column percentages—the proportion of latent class choice model segment respondents correctly classified. One can see there is strong agreement between the two methods, 84.5% overall. This is less than we saw in the earlier data set. The "low scale" class (#2) shows only a 76.7% agreement. This may be a result of 2 scale classes not being the optimal number of scale classes, and of the degree of heterogeneity across the 5-preference segments.

Table B4: Scale Class Membership Cross Tabulation

		Best-Worst Response Scale Classes **Col Pct**	
		Scale class 1	Scale class 2
HB Utilities Scale Classes	Scale class 1	1979 88.7%	280 23.3%
	Scale class 2	251 11.3%	923 76.7%

Table B5 shows that 68.2% of the respondents were put into corresponding preference segments. This is considerably higher than if we had not used scale classes. Again, the rows represent the SALC clustering of the HB utilities using 2 scale classes and the columns represent the SALC choice model. Both are run with the choice and non-choice data included. The segments do not match in their numbering. That is, preference segment 1 of the SALC choice segmentation is the equivalent of preference segment 3 of the SALC clustering of the HB utilities. The cells highlighted in tan/orange are the matched segments. Among the cells of agreement, the overall match is 68.2%. There are several cells shaded in lighter gray, highlighting situations where over 100 respondents are classified into segments that did not match the LC choice model segments.

Table B5: Preference Segment Cross Tabulation



		Best-Worst Response Preference Segments Col Pct				
		Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
HB Utilities Preference Segments	Segment 1	61 9.6%	101 16.0%	22 3.6%	599 65.7%	31 4.8%
	Segment 2	9 1.4%	398 63.2%	21 3.4%	237 26.0%	82 12.8%
	Segment 3	440 69.3%	13 2.1%	120 19.5%	8 0.9%	41 6.4%
	Segment 4	79 12.4%	16 2.5%	437 70.9%	14 1.5%	80 12.5%
	Segment 5	46 7.2%	102 16.2%	16 2.6%	54 5.9%	406 63.4%

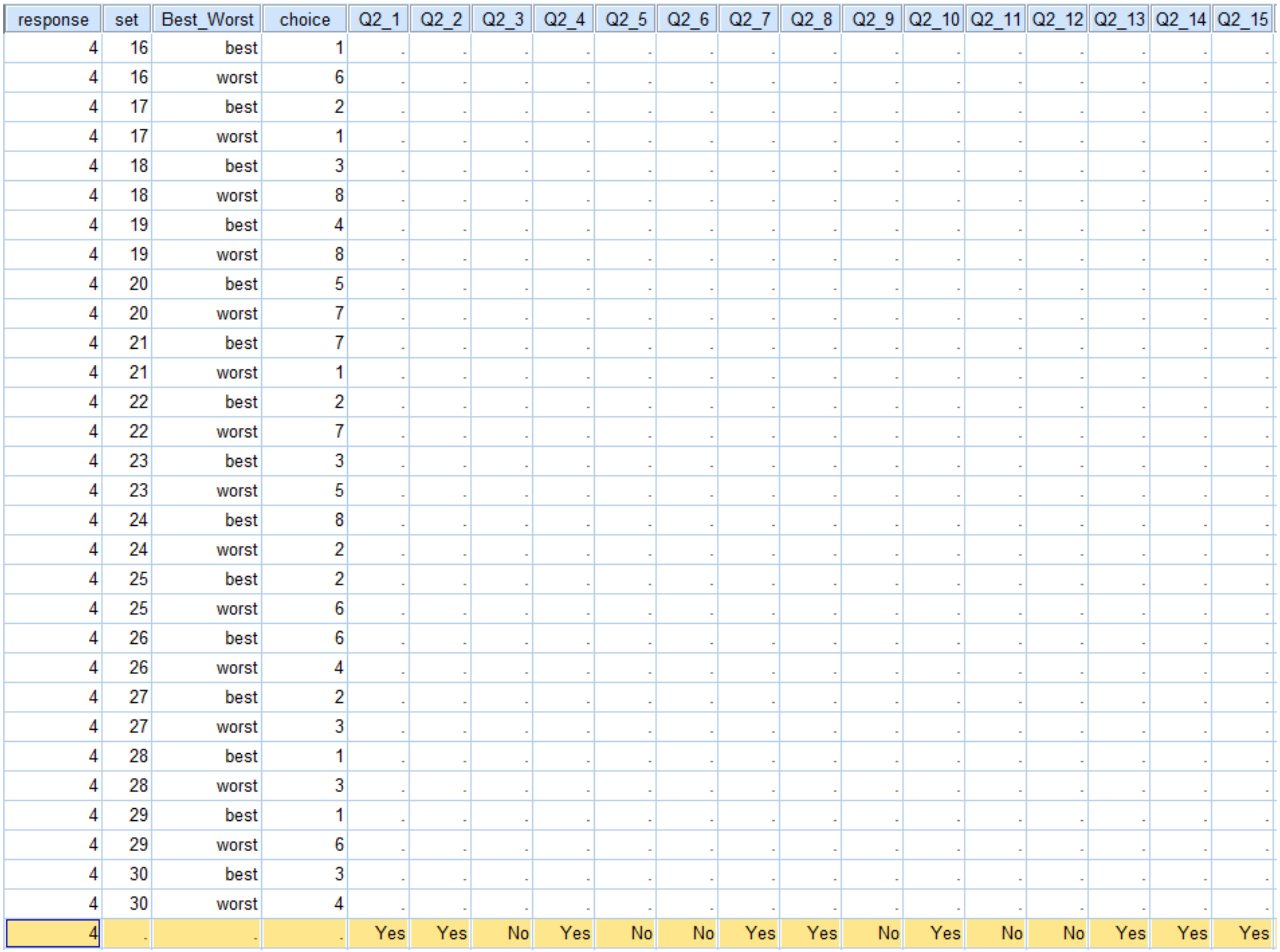
APPENDIX C: BEST-WORST (MAXDIFF) DATA FILE FORMAT FOR LATENT GOLD® WITH ADDITIONAL NON-CHOICE VARIABLES

Latent Class segmentation based on choice or non-choice response data can be conducted using the Latent GOLD® program (see e.g., https://www.statisticalinnovations.com/latent-gold-5-1/). This Appendix illustrates and discusses the data format required by Latent GOLD in the case where the response data consists of *both* 1) Best and Worst (MaxDiff) choices as well as 2) responses to non-choice variables to be used as additional basis variables. Table C1 illustrates the Latent GOLD setup[13] using the Best-Worst data from the Australian Health Reform Study, described in the main body of this paper, when combined with the additional 15 dichotomous responses Q2_1–Q2_15, elicited from the following survey question:

> "Please tick those principles that you think should be considered by the government in terms of Health Care Reform for each of the 15 principles" [listed earlier in Table 2].

[13] Latent GOLD allows for two different formatting options—the 3-file format, and the 1-file format. For simplicity, we only illustrate the 3-file format here. For further details, including the corresponding illustration for the 1-file format, see Vermunt and Magidson (2011).

Table C1. Illustration of the Response File in Latent GOLD's 3-file format for respondent #4.



response	set	Best_Worst	choice	Q2_1	Q2_2	Q2_3	Q2_4	Q2_5	Q2_6	Q2_7	Q2_8	Q2_9	Q2_10	Q2_11	Q2_12	Q2_13	Q2_14	Q2_15
4	16	best	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	16	worst	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	17	best	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	17	worst	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	18	best	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	18	worst	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	19	best	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	19	worst	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	20	best	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	20	worst	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	21	best	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	21	worst	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	22	best	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	22	worst	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	23	best	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	23	worst	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	24	best	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	24	worst	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	25	best	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	25	worst	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	26	best	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	26	worst	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	27	best	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	27	worst	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	28	best	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	28	worst	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	29	best	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	29	worst	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	30	best	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	30	worst	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	Yes	Yes	No	Yes	No	No	Yes	Yes	No	Yes	No	No	Yes	Yes	Yes

Specifically, the first 30 records included in Table C1 illustrate the choice response data for respondent #4—the best and worst choices for this respondent. The final (31st) record (highlighted in Table C1) contains the responses to the non-choice questions. The complete response file contains similar records for each of the other 203 respondents and stacked together in this file. In the *standard* setup for estimating a LC Best-Worst model (without responses to the non-choice variables), the single non-choice record would be omitted for each respondent.

Appendix D: Syntax of SALC Model for HB Utilities for Both Choice and Non-Choice Data with and without Differential Variable Weighting

Latent GOLD 6.0 Syntax *Without* Differential Variable Weighting

```
variables
caseid respid;
dependent (asc_1 – asc_14) continuous,
(nq19_1 – nq19_11) ordinal; // 14 HB vars are asc_1 – asc_14; 11 attitudinal vars are nq19_1 – nq19_11
latent
Cluster nominal 5, sclass nominal 2 coding = 1;        // 5 preference classes and 2 scale classes
equations
Cluster <- 1; sclass <- 1 ;        // size parameters for preference classes and scale classes
asc_1 – asc_14 nq19_1 – nq19_11 <- 1 + Cluster;
asc_1 – asc_14;
asc_1 – asc_14 <<- (s) sclass;// Scale adjustment is applied only to the HB choice variables
s=-;
```

Latent GOLD 6.0 Syntax with HB Variables *Down-Weighted* by .5

```
variables
caseid respid;
dependent (asc_1 – asc_14) continuous varweight = .5,
(nq19_1 – nq19_11) ordinal varweight = 1.0;
latent
Cluster nominal 5, sclass nominal 2 coding = 1;
equations
Cluster <- 1; sclass <- 1 ;
asc_1 – asc_14 nq19_1 – nq19_11 <- 1 + Cluster;
asc_1 – asc_14;
asc_1 – asc_14 <<- (s) sclass;
s=-;
```

References

Eagle, T. (2013). "Segmenting Choice and Non-Choice Data Simultaneously," *2013 Sawtooth Software Conference Proceedings*, pp. 231 –250.

Eagle, T., J. Louviere, and T. Islam, (2018) "A Comparison of Volumetric Models," *2018 Sawtooth Software Conference Proceedings*, pp 267–292.

Groothuis-Oudshoorn, C.G.M., T.N. Flynn, H.L. Yoo, J. Magidson and M. Oppe (2018). "Key Issues and Potential Solutions for Understanding Health Care Preference Heterogeneity Free from Patient Level Scale Confounds," *The Patient: Patient-Centered Outcomes Research*, https://rdcu.be/Mx8e

Hess, S., and J.M. Rose (2012). Can scale and coefficient heterogeneity be separated in random coefficients models? *Transportation*, 39, 4.

Lee, J. and J. D. Brazell (2019). "Multivariate analysis with MaxDiff: A curious look at the properties of MaxDiff utilities," AMA Advanced Research Techniques Forum, Brigham Young University, Provo, UT.

Lindsay, B., C. C. Clogg, and J. Grego (1991). "Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis," *Journal* of the *American Statistical Association*, Vol. 86, No. 413, pp. 96–107.

Louviere, Jordan J., and Thomas C. Eagle, (2006). "Confound it! That Pesky Little Scale Constant Messes Up," *2006 Sawtooth Software Conference Proceedings*.

Louviere, J. J., and T.N. Flynn (2010). "Using best-worst choice experiments to measure public perceptions and preferences for health care reform in Australia," *The Patient: Patient-Centered Outcomes Research, 3*(4), 275–283.

Lyon, D. W. (2020). "Comments on 'Segmenting Choice and Non-Choice Data Simultaneously: Part Deux,'" *2019 Sawtooth Software Conference Proceedings*.

Magidson, J., D. Thomas, and J.K. Vermunt (2009). A new model for the fusion of MaxDiff scaling and ratings data. *2009 Sawtooth Software Conference Proceedings*, 83–103.

Magidson, J., and J.K. Vermunt (2001). Latent class factor and cluster models, bi-plots and related graphical displays. Sociological Methodology, 31, 223–264.

Magidson, J., and J.K. Vermunt, (2007). "Removing the Scale Factor Confound in Multinomial Logit Choice Models to Obtain Better Estimates of Preference," *2007 Sawtooth Software Conference Proceedings*, pp 139–154.

Vermunt, J.K., and J. Magidson (2004). Factor analysis with categorical indicators: a comparison between traditional and latent class approaches. In: Van der Ark, A., Croon, M.A., and Sijtsma, K. (Eds), *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*. Erlbaum.

Vermunt J.K., and J. Magidson (2013). *Latent GOLD 5.0 upgrade manual*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J.K., and J. Magidson (2011). *LG-Syntax User's Guide: LG Choice Modeling Extensions via the LG-Syntax Module*, Belmont, MA: Statistical Innovations Inc.

COMMENTS ON "SEGMENTING CHOICE AND NON-CHOICE DATA SIMULTANEOUSLY: PART DEUX"

DAVID W. LYON
AURORA MARKET MODELING, LLC

INNOVATIONS IN EAGLE AND MAGIDSON

The preceding paper by Tom Eagle and Jay Magidson introduces two important technical innovations in latent class modeling, while reinforcing an earlier one (non-continuous SALC). It illustrates their application with very well-chosen case studies.

Their analysis of the Flynn health-care data is a particularly convincing validation study. Going beyond the usual "here's what we did, here's what we got" case study, it brings in substantive subject-matter expectations, and demonstrates how scale adjustment confirms those expectations, while classic or "straight" LC modeling does not. That turns out to be true both for their scale-adjusted choice model-based LC, and for LC applied to the HB utilities, using discrete and continuous scale adjustment, respectively. Such validation is particularly impressive because scale-adjusted latent class (SALC) is a purely technical feature, not in any way tailored to the specific hypothesis or subject area on which it was tested.

Entirely aside from use with HB utilities, continuous-variable SALC will be useful with any continuous variables, just as the original categorical-variable SALC works with any nominal variables.

The travel case study illustrates how differential variable weighting can be used to control and "tune" segmentation results. As the authors point out, this capability will take practitioners some time to learn to use judiciously. But it obviates a host of clunky ways practitioners have attempted to reach the same goal—things like using factor scores to reduce the influence of a group of variables, or picking some to simply drop, or increasing influence by entering some variables twice. This new capability is like handing a nice rubber mallet to a sculptor who has had only a sledgehammer.

In addition to the technical aspects, Eagle and Magidson remind us once again that segmenting on both choice and non-choice data is not only possible but often very useful. The substantive needs of real-world segmentation often demand using both kinds of data. Technical complexities or data file setup issues are minor in the overall scheme of things, and not excuses to oversimplify segmentation needs.

THE GOLD STANDARD

The authors mention several times that, with choice data, segmenting on the choice model or choice data itself (as opposed to segmenting on HB utilities estimated from the choice data) is the "gold standard" approach. This echoes Eagle (2013; full citation at end of Eagle and Magidson paper), a paper that was emphatic on that point. They also show that, at least in their two case studies, SALC on HB utilities (zero-centered) produces segments in close agreement (85% or more) with those produced by SALC on the choices themselves.

The primary objective of these comments is to:

- Examine and illustrate *why* latent class on choice data (not utilities) is the gold standard.
- Argue that using SALC on utilities does *not* necessarily overcome the problems of using utilities in latent class.

Consider the segment agreement tables shown just below. Each shows two different 3-segment solutions, with each pair agreeing barely 40% of the time. With 33% being the minimum possible agreement for 3-segment solutions, they all show exceptionally poor agreement among different solutions.

38	60	7
18	**47**	5
15	12	**2**
	43%	

61	28	15
47	**20**	6
11	14	**2**
	41%	

49	63	18
8	**31**	7
12	13	**3**
	41%	

60	33	32
37	**13**	12
7	4	**6**
	39%	

What makes such poor results relevant here is that all 8 of those solutions were obtained by running Latent Gold (using standard LC cluster, not SALC) *on the same HB utilities*! Specifically, they were run on the HB utilities Tom Eagle estimated from the Flynn data,[1] the ones Eagle and Magidson used in their first case study.

How Utilities Work

How is this possible? It is not because Latent Gold is defective. The input it was given for these 8 solutions were the "same utilities" in every case, but not actually the same *numbers*. The utilities were simply zero-referenced in different ways for each solution.

Recall that utilities are derived from a multinomial logit equation,

$$\boldsymbol{p(a|A)} = \frac{\mathbf{ex\,p}(\boldsymbol{U_a})}{\sum_{\boldsymbol{j \in A}} \mathbf{ex\,p}(\boldsymbol{U_j})}$$

In this equation, adding any constant, C, to every item's utility, changes nothing because the added constant cancels out:

$$\boldsymbol{p(a|A)} = \frac{\boldsymbol{exp(U_a + C)}}{\sum_{\boldsymbol{j \in A}} \boldsymbol{exp(U_j + C)}} = \frac{\boldsymbol{exp(C)\ exp(U_a)}}{\boldsymbol{exp(C)}\ \sum_{\boldsymbol{j \in A}} \boldsymbol{exp(U_j)}} = \frac{\boldsymbol{exp(U_a)}}{\sum_{\boldsymbol{j \in A}} \boldsymbol{exp(U_j)}}$$

This means utilities are "unidentified" until we decide how to pin them down. Most commonly, we zero-center them, so their overall average is zero. It is also common to pick one item, often the last, to fix at a zero utility. But there are infinitely many other possibilities as well. The 8 disagreeing solutions shown above simply reflect solutions from zero-centered utilities and from 7 different choices of which item to fix at a zero utility.

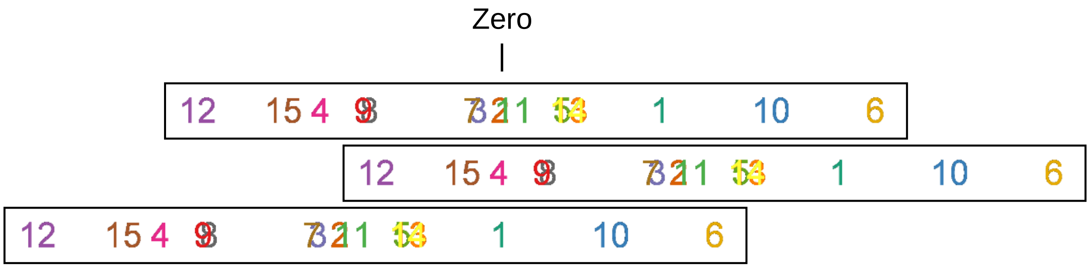
Note that how we "center" or "reference" the utilities has no impact whatever on how well they fit or describe the choice data and no impact whatever on what they predict or

[1] I thank Tom Eagle for providing a file of this data, along with the authors' scale-adjusted HB-utility–based latent class solution. Indirectly, I join the authors in thanking Terry Flynn for making this data available.

would simulate for any future choice set.[2] In every substantive sense, the utilities are truly "the same" regardless of how zero-referencing is done.

In aggregate models, the choice of reference has essentially no impact at all. But in individual-level (e.g., HB) data, the referencing operates differently for each individual. Latent class results, and utility correlations, and any other analyses that depend on patterns of utilities *across* individuals all turn out differently depending on how the utilities are zero-referenced.

We can think of each respondent's utilities as being fixed on a wooden ruler that shows item numbers positioned to reflect that respondent's utilities. On such a ruler, the spacing of utilities and their absolute differences are meaningful,[3] but which point we choose to call zero is not (picking the zero point is equivalent to choosing the additive constant C in the MNL equation shown earlier). Here is the "ruler" for the fourth respondent in the Flynn file, positioned once with zero-centering, once with item 4's utility set to zero, and shifted again to set item 1's utility to zero.



The non-identification of the utilities means we can slide the rulers left and right as far as we like, relative to the zero point, without changing their real meaning.

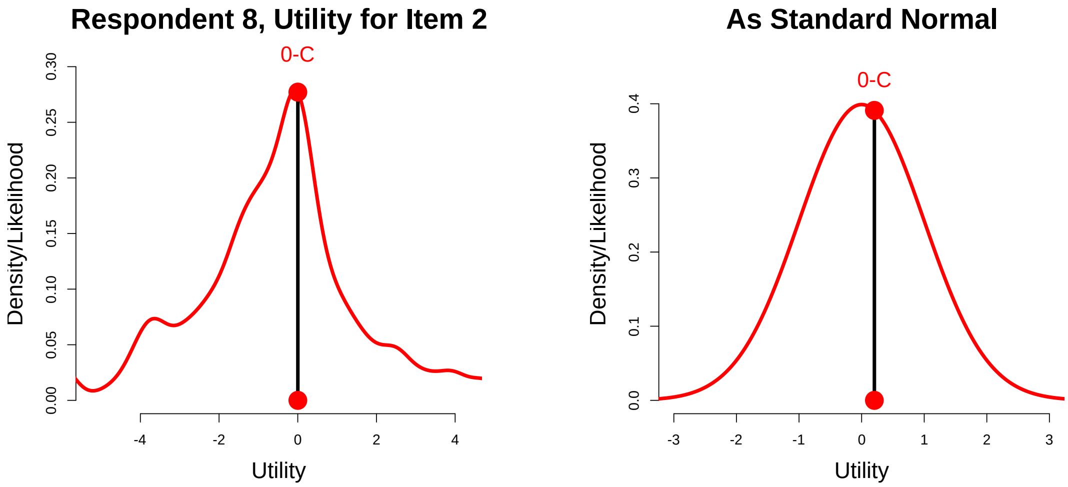
Likelihoods for Utilities

When we segment on HB utilities, latent class modeling is based on the likelihood of a given respondent's utility on a particular item, given the distribution of that item's utility across all respondents in a segment. Most software bases that likelihood on a normal approximation to the actual distribution. Using zero-centered utilities, the 4th Flynn respondent (ID number 8) has a zero on utility 2, and his or her utility falls in the middle of the distribution across all respondents, as illustrated in the two graphs below.[4]

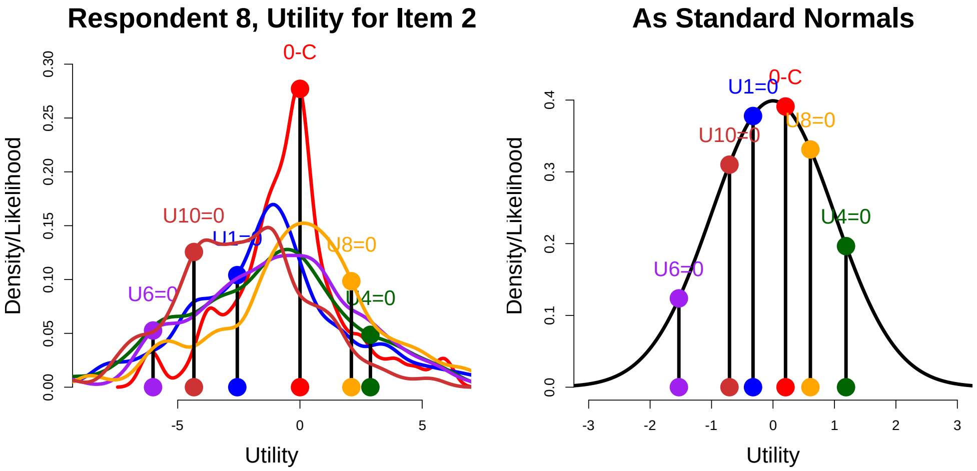
[2] This is true only for additive "centering." Multiplying utilities by a constant does change their fit and their predictions, as does any non-linear transformation. The popular "zero-centered diffs" rescaling involves multiplying by a different constant for each respondent, so it does not preserve the predictive properties of the MNL.

[3] They are "meaningful" in the sense that they determine predicted choices, and predict differently when changed. Some argue they are not meaningful from the viewpoint of utilities being a combination of preferences, which are regarded as meaningful, and error scale, which is not. The key issue in our context is the impact on predictions, which is not to dispute the other viewpoint in contexts focused on interpretation.

[4] This is illustrated here based on the total sample distribution; what actually matters as latent class iterates toward a final solution is the distribution within each separate latent class. The principle is the same.

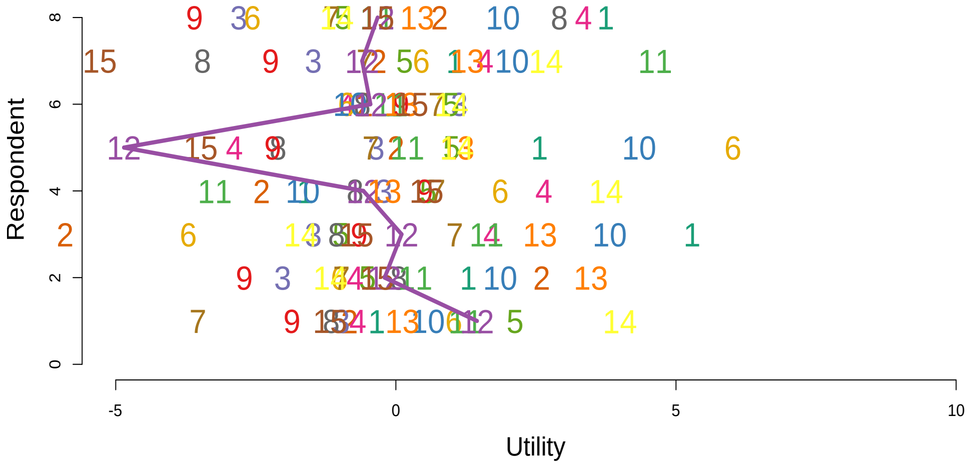


But what happens if we use a different reference point? The next two graphs show the results for six different referencing options. Is the respondent right in the middle of the distribution, with a high likelihood, or off at the side (as with utility 6 being set to zero) with a much lower likelihood? We can't even be sure if this respondent is relatively low (utility 6 set to zero) or high (utility 4 being zero) relative to the other respondents, let alone exactly how high the likelihood is!

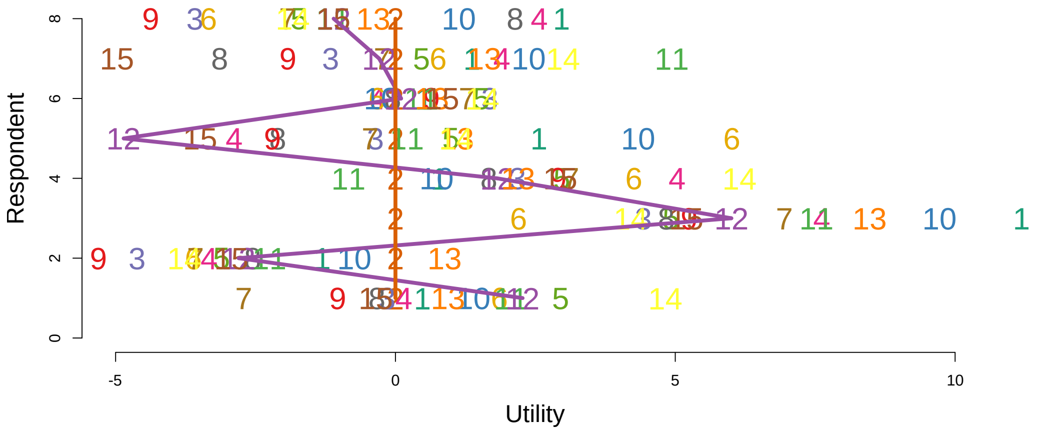


How can this happen? The problem is that each individual's utility "ruler" is adjusted by a different amount when we choose a new reference point. If we simply added the identical constant C to all utilities for *all* respondents, nothing would change in those graphs except the absolute numbers on the x-axes. But with every respondent being independently zero-centered, or item 4 zero-referenced, say, each one moves in a different way. Again, the fit and predictions don't change, but the relationship of respondents *to each other* shifts around.

Consider how the pattern of utilities for item 12 shifts with changes in reference. The next chart shows the utility rulers for the first 8 Flynn respondents, with the line connecting their utilities for item 12, on a zero-centered basis.



If, instead of zero-centering, we set the utility of item 2 to zero for everyone,[5] that pattern shifts dramatically:



So, any segmentation results based on utilities are very dependent on a fundamentally arbitrary decision as to how to uniquely identify the utilities. What can we do?

Likelihoods for Choices

The solution is simply to use the actual choice data for latent class, in the context of a latent class *choice* model. In an LC choice model, the likelihood does *not* depend on a respondent's position relative to others, but is the probability of the respondent's actual choice in one task, given the utility for that segment.

[5] Is this a silly choice of reference made just for arguments' sake? Not necessarily! Suppose item 2 were "not change anything in our product" and the other 14 were various kinds of changes, variously liked or disliked by different respondents. This would then be a wholly natural way of referencing, and far more readily interpretable than any other.

No matter how the utilities are coded in the program, the *likelihood* of a given choice stays the same. This is true because *any* referencing or coding scheme produces the same predicted choices. The software must still make an arbitrary choice of coding scheme, but its choice has no impact on the *predictions* and thus no impact on the likelihoods.

So, why is using the choice data directly, in a latent class choice model, the gold standard? Because it requires no arbitrary, results-changing, decisions! It is invariant to things that don't matter, unlike LC clustering on HB utilities.

IS ZERO-CENTERING SPECIAL?

It is important to understand that zero-centering utilities does not escape the general problem. While zero-centering is very widely used, and often feels very natural, it is still just one of the infinitely many ways utilities could have been identified. Choosing it is just as arbitrary as any other choice.

Zero-centering does tend to minimize correlations (across respondents) among utilities, and minimizes the overall variance of all utilities taken as a group. It is possible, but unproven, that this could in turn cause clusters based on them to more closely match segments based on choice data. Eagle and Magidson do get good agreement between the two in their two case studies (using scale adjustment in each case) and have told me in discussions that they believe this is likely a general effect. But, much more investigation will be needed to see whether it is a general finding. Even if it is true in general, why use it? Latent Gold and Sawtooth Software products both offer latent class *choice* solutions. They are truly the gold standard and should be used.

Zero-centered diffs, also often used, offer no escape either; they only compound the issue. Because they use multiplicative scaling on the utilities, they destroy the link to predicted probabilities, in addition to beginning with an arbitrary choice of referencing.

A reviewer of these comments posed an interesting hypothetical: If you were given utilities, and had no access to the original choice data, wouldn't you go ahead and cluster on the utilities (zero-centered, let's say)? My answer is no, I would not.[6] Instead, I would generate a large[7] choice design, simulate answers to it from the utilities plus Gumbel error and then run latent class choice models on the simulated answers.

IS THIS A PROBLEM WITH LATENT CLASS? OR WITH HIERARCHICAL BAYES?

It's also important to note that none of the problems discussed here are unique to latent class clustering. LC clustering is the most popular way of finding segments in market research today, for many well-known reasons. As the fundamental topic of the Eagle and Magidson paper that inspired these comments, it is the obvious context in which to embed this discussion. But the problem of basing segmentations on arbitrarily-referenced utilities applies just as well to *any* clustering approach, including K-means, hierarchical methods, ensemble methods, and all the many others. It is not in any way specific to LC clustering. The problem is in the utilities, not in the clustering method.

[6] Unless it were all due in the next 4 hours; I am a practitioner, after all.

[7] "Large" meaning *much* larger than typically used with real respondents. We want the effects of the Gumbel error to average out, and there is no cost to using more simulated tasks, so we can use 10 or 20 times as many as in a real-world MaxDiff design.

Indeed, for choice data, LC is clearly superior to clustering approaches because it does offer the LC choice modeling approach as an alternative to LC clustering on utilities. No other clustering method known to the author can deal directly with choices in a reasonable way; one has no option but to use utilities with the other methods and suffer all the problems discussed here.

Similarly, this is not a hierarchical Bayes (HB) problem. HB is by far the most common method in use for obtaining individual-level utility estimates, although not the only one. Again, the problem is not in the method of estimation, but in the fact that utilities, by their very definition in the MNL, inherently require an arbitrary referencing decision. That decision affects their distribution across respondents, which is what ultimately drives any clustering method.

Conclusions

The bottom line is that with any type of choice data, latent class modeling should be done using latent class choice models, not using latent class clustering on individual-level utilities from an initial HB analysis (or any other method of calculating individual utilities, for that matter). If utilities are used, results are heavily dependent on an arbitrary identification decision.

Eagle and Magidson show good agreement between segments from the choice data and segments from the HB utilities (which were zero-centered ones in their analysis). It is a safe bet that other referencing points would have produced much worse agreement. Some may read their paper to condone latent class clustering on HB utilities (despite their explicit advice against it), which would be unfortunate. SALC on continuous variables has many other potential applications; to think that it "rescues" utility-based segmentations in general is a mistake. Applied to zero-centered utilities, it *might* produce results closely resembling the gold standard LC choice models, but the evidence for that is limited so far, and it is hard to see from first principles why that should be.

We should acknowledge that many practitioners, including this author, as well as Eagle and Magidson, have used latent class segments based on utilities, and obtained meaningful and useful segments. We've all done it, it seems to work just fine! That does not counter the argument against doing so. Segmentation is, in many ways, an easy problem—almost any way of dividing up a sample, based on almost any variables of substantive relevance, will produce segments that are far more meaningful and useful than treating the world as a single mass market. Even consultants eyeballing short questionnaires from focus group participants, and putting them in separate piles by pure judgment, have created usefully different groups.

The relative ease of success in segmentation does not make such methods principled, let alone optimal in any meaningful sense. When working with choice-based methods, we have the machinery (latent class choice models, and SALC choice models) to do far better, and we should! We might just discover how much better we could have been doing all along.

With scale-adjusted latent class methods no longer restricted to categorical variables, and with differential weighting to control the combination of choice and non-choice data (among other uses), Eagle and Magidson have made the available machinery better yet.[8]

David W. Lyon

[8] I thank Jay Magidson, Tom Eagle, Bryan Orme, and Keith Chrzan for their reactions to earlier drafts of these comments. All led to improvements, though almost all disagreed at least in part with the views expressed here. All remaining errors and misconceptions are the author's alone!

UNDERSTANDING CONSUMER PREFERENCES: A COMPARISON OF SURVEY- AND PURCHASE-BASED APPROACHES

JAMES PITCHER
BRADLEY TAYLOR
DAN KELLY
GFK

ABSTRACT

Can Point of Sale (POS) data tell us anything about consumer preferences? We conducted one of the largest and most comprehensive research studies to calculate attribute importance, brand preference, and price elasticities from POS data, and compared the results with those obtained from a CBC exercise across 15 technology and durables product categories within 7 countries; 68 cells in total. We see that although attribute importance and brand preferences are similar, there are large differences in price elasticities between the POS and conjoint models due to the differing ways in which they measure consumer preferences. Conjoint measures a theoretical preference, one that is not influenced by external market factors, whereas the POS data takes into account the in-store realities and how these affect the purchase decision.

BACKGROUND AND DESCRIPTION

Conjoint analysis is the gold standard methodology for measuring consumer preferences. However, survey-based approaches face increasing pressure to maintain engagement with respondents living ever-busier lives and to deliver insights to clients faster and at a reduced cost. For these reasons, steps are being taken to streamline surveys and harness the power of imputation methods in order to reduce survey length. But what if we consider an alternative solution where we replace parts of a survey with behavioural data? Instead of measuring what people say they will do, we measure what they actually did.

At GfK, we have access to Point of Sale (POS) data that contains a huge amount of information on product features, prices, and product sales. Over 760 product groups are audited globally, with 1.5M new products being added each year. The data covers more than 120 channels of distribution and over 425,000 stores worldwide. It means we have a rich amount of store-level data on individual products. For each individual product we have the following information:

- Product specification and features (e.g., For a Laptop: brand, screen size, storage capacity, CPU, RAM, battery length, etc.).
- Units sold and revenue generated in each store per week.
- The price the product was sold at in each store per week.
- What price discounts (if any) were applied in each store per week.
- Share of shelf in each store per week.
- Distribution: availability of the product in each store per week.

Can we use this rich data source to tell us anything about consumer preferences? Specifically, can we measure the appeal of different brands and product features, price sensitivity, and how important brand, price and features are in the purchase decision? How do these results compare to those obtained through a traditional conjoint exercise? Where do they complement each other and where do they show us something different or interesting? What are the limitations of the POS data?

We directly compare how POS data and a traditional conjoint exercise predict consumer preferences by running a parallel study where we use both approaches to determine the attractiveness of product features and price sensitivities across multiple technology and durables product categories.

STUDY DESIGN

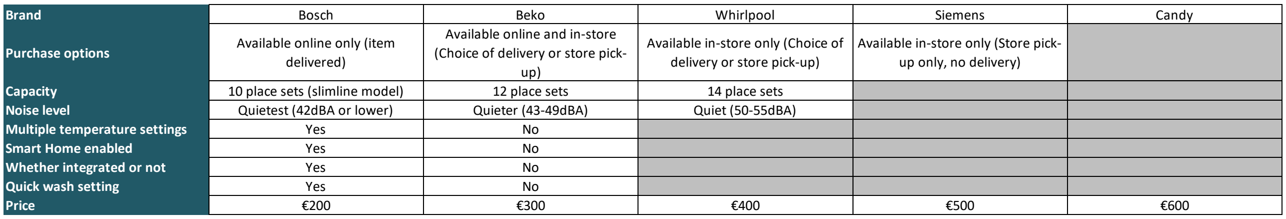
The study was run across 15 technology and durables product categories within 7 countries, 68 cells in total. Both conjoint and POS analysis was conducted in each cell.

Product Categories	**Brazil**	**France**	**Germany**	**Great Britain**	**Russia**	**Japan**	**China**	Grand Total
Smart Speakers	[redacted]	1	1	1	[redacted]	[redacted]	[redacted]	3
Cooking (Oven)	1	[redacted]	[redacted]	[redacted]	1	[redacted]	[redacted]	2
Dishwashers	[redacted]	1	1	1	[redacted]	[redacted]	[redacted]	3
Refrigerators	1	1	1	1	1	[redacted]	[redacted]	5
Digital Cameras	[redacted]	[redacted]	1	[redacted]	1	[redacted]	[redacted]	2
Hot Beverage Makers	1	1	1	1	1	1	1	7
Irons	1	[redacted]	1	[redacted]	1	[redacted]	[redacted]	3
Tablets	1	1	1	1	1	[redacted]	[redacted]	5
Mobile Computing (Laptop/Notebooks)	1	1	1	1	1	1	1	7
TV (PTV/ Flat (LCD/ Plasma/ Rear))	1	1	1	1	1	1	1	7
Shavers (Men’s)	1	[redacted]	1	[redacted]	1	[redacted]	[redacted]	3
Ladies Epilators/Laser IPL	1	[redacted]	1	[redacted]	1	[redacted]	[redacted]	3
Smartphones	1	1	1	1	1	1	1	7
Vacuum Cleaners	[redacted]	1	1	1	1	[redacted]	[redacted]	4
Washing Machines	1	1	1	1	1	1	1	7
Total	**12**	**10**	**14**	**10**	**13**	**5**	**5**	**68**

Conjoint

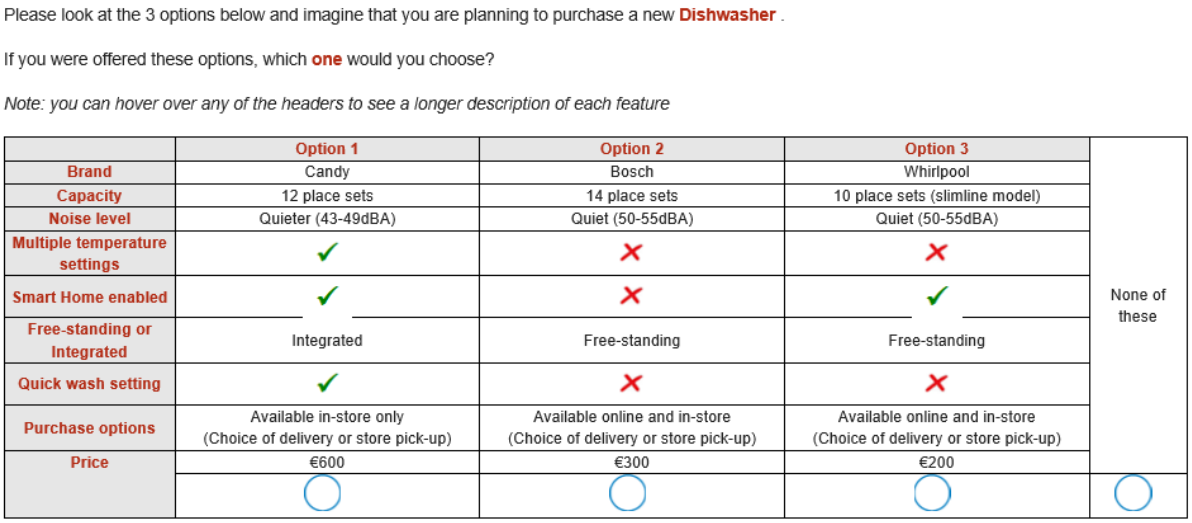
Within each cell, a representative sample of 500 respondents answered a Choice-Based Conjoint (CBC) exercise with 12 tasks. We interviewed around 38,000 respondents in total. Two designs were created, consisting of nine attributes: brand, price, purchase options, and 6 attributes relating to product features. 3 to 4 were on/off attributes. Fieldwork was conducted in March 2019. Part-worth utilities for all levels were estimated in Sawtooth's Lighthouse studio.

Attribute Grid for Dishwashers in France



Brand	Bosch	Beko	Whirlpool	Siemens	Candy
Purchase options	Available online only (item delivered)	Available online and in-store (Choice of delivery or store pick-up)	Available in-store only (Choice of delivery or store pick-up)	Available in-store only (Store pick-up only, no delivery)	
Capacity	10 place sets (slimline model)	12 place sets	14 place sets		
Noise level	Quietest (42dBA or lower)	Quieter (43-49dBA)	Quiet (50-55dBA)		
Multiple temperature settings	Yes	No			
Smart Home enabled	Yes	No			
Whether integrated or not	Yes	No			
Quick wash setting	Yes	No			
Price	€200	€300	€400	€500	€600

Conjoint Tasks for Dishwashers in France



Please look at the 3 options below and imagine that you are planning to purchase a new **Dishwasher** .

If you were offered these options, which **one** would you choose?

Note: you can hover over any of the headers to see a longer description of each feature

	Option 1	Option 2	Option 3	
Brand	Candy	Bosch	Whirlpool	None of these
Capacity	12 place sets	14 place sets	10 place sets (slimline model)	
Noise level	Quieter (43-49dBA)	Quiet (50-55dBA)	Quiet (50-55dBA)	
Multiple temperature settings	✓	✗	✗	
Smart Home enabled	✓	✗	✓	
Free-standing or Integrated	Integrated	Free-standing	Free-standing	
Quick wash setting	✓	✗	✗	
Purchase options	Available in-store only (Choice of delivery or store pick-up)	Available online and in-store (Choice of delivery or store pick-up)	Available online and in-store (Choice of delivery or store pick-up)	
Price	€600	€300	€200	
	○	○	○	○

POS Models

For simplicity, we only considered offline sales. Online sales were not included. To calculate attribute importance and brand preferences, we used 18 months' worth of weekly store-level data between October 2018 and April 2019. When calculating product price elasticities, we used a reduced modelling period of six months, between February 2019 and June 2019. This was to reduce any bias caused by product lifecycles and changes to the competitive landscape.

Analysis

We calculated Attribute Importance, Brand Preference, Feature Preference/Pricing, and Price Elasticities from both the conjoint and POS data. In this section, we detail how these measures were calculated and compared.

Conjoint Attribute Importance

Conjoint attribute importance scores were calculated using the standard method: using the zero-centred diffs utility scores, we calculated the range in utility scores for each attribute and calculated importance values that sum to 100 percent. The importance scores of all attributes relating to features were summed to give an overall importance score for features as whole.

POS Attribute Importance

For each category and country, we created a Random Forest Model, using the "h2o.randomForest" function in the "H20" package in R, to predict sales units within each retailer using brand, features, and price, while controlling for seasonality, retailer distribution, and price discounts. We used the top five selling brands and up to 40 product features, depending on the availability in the category. We made the distinction between the product base price and price discount to ensure that we calculated the importance of everyday price, not large promotional price reductions. A temporary price reduction is defined as a price reduction higher than 10% that lasts more than six weeks. The base price was then split into five equal bins. Seasonality was included as a control term by calculating the average total sales of the category per week. Distribution was also included as a control variable and was recoded to be on a scale of 0 to 1 and divided into buckets of size 0.05. The data was mean-centred at the retailer level to remove any scale affect resulting from the difference in volumes sold from store to store.

The importance of brand, price, and features was determined by calculating the relative influence of each variable. More specifically, whether that variable was selected to split on during the tree building process, and how much the squared error (over all trees) improved (decreased) as a result (H2O.ai, 2019). The importance scores of all attributes relating to features were again summed to give an overall importance score for features as whole.

Conjoint Brand Preference

The relative appeal of each brand was assessed by calculating the average zero-centred diffs utility scores for each brand.

POS Brand Preference

We calculated a score for each brand by comparing how the actual sales of a brand differed from the expected sales for that brand, given the brand's average price and distribution compared to the category as a whole:

- Observed Share = Total Brand Unit Share
- Expected Share = (Brand Distribution/Category Distribution) * (Average Price of Brand/Average Price of Category)
- Brand Score = Observed Share/Expected Share

The higher the observed sales, compared to what was expected, the higher the strength of the brand.

Conjoint Feature Preference

The relative appeal of each feature was assessed by calculating the average zero-centred diffs utility scores for each feature.

POS Feature Pricing

We were unable to calculate consumer preferences using the POS data due to the high level of multicollinearity between product features. Features do not appear in enough unique combinations across products for us to isolate the appeal of each individual feature.

Instead, we calculated the price that manufacturers charge for specific features. We take the average price of all products with a given feature and compare it with the category average price to calculate the premium or discount for that feature. Therefore, for each feature we have a complete overview as to whether the presence (or absence) of a given feature carries a premium versus the category average. Note that this is not a measure of consumer preferences as it is not related to sales. It is simply how much extra retailers charge for a feature.

Conjoint Price Elasticities

We first created a scenario consisting of the 20 top selling products as observed in the POS data. However, if there was insufficient POS data to model the product price elasticity, the product was excluded from the conjoint base scenario. We then calculated the share of preference of each product, one at a time, at five prices between the minimum and maximum prices observed in the POS data for the given product. The price of the other products was held constant. The price elasticity was calculated by taking the natural log of five prices and resulting shares and finding the regression slope that best fits the data.

POS Price Elasticities

We consider the 20 top-selling products as observed in the POS data. However, there was insufficient data to model the product price elasticity for some products. For each product, we created a separate multiplicative regression model at the store level to predict sales units using the product’s own base price, own price discount, the base price and price discounts of competitors, and presence of competitors in store. We made the distinction between the base price and price discounts to ensure that we calculated the impact of everyday price, not large promotional price reductions. We also controlled for category seasonality and trend using a LOESS smoothing algorithm. Sales units and all base prices were log transformed. The data was mean-centred at the store level to remove any scale affect resulting from the difference in volumes sold from store to store. The coefficient for the product’s base price represents the price elasticity of the product.

Price elasticities were only calculated across seven product categories because it was a very labour-intensive process to run the POS models since a separate manually-fitted model was required for each product. Price elasticities were calculated for 38 products in total. Category elasticities were calculated by taking the average price elasticity for all products within the given category.

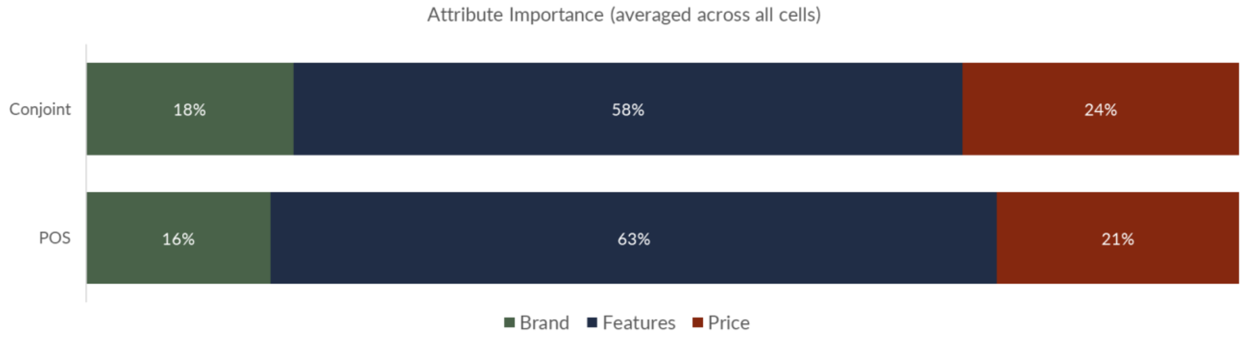
RESULTS

Although we started with 68 cells, in many of the cells we were unable to use the POS data to produce sufficiently robust models. This was due to having either too little sales data or too much week-on-week variation in the data. As a result, most results are based on 38 cells.

Attribute Importance

Figure 1 shows that when the importance scores are averaged across the 38 cells that were analysed, the attribute importance scores derived from the conjoint and POS data are remarkably consistent. Features make up around 60% of the purchase decision in both the conjoint (58%) and the POS model (65%). Features rank as the most important attribute in all 38 cells in the conjoint, and in 35 cells (92%) in the POS data.

Figure 1: Attribute importance scores averaged across all cells.



However, when we look within each individual cell, we see differences in importance scores. Figure 2 shows the absolute differences in attribute importance (POS - Conjoint) in each individual cell. In some cells, such as Hot Beverage Makers in Japan, we observe very small differences (Figure 3). But we consistently see the that feature importance is much higher for Laptops and Tablets in the POS model compared to the conjoint. Figure 4 shows feature importance in UK Tablets is almost double in the POS compared to the conjoint.

Figure 2: Absolute differences in attribute importance (POS - Conjoint) in each cell.

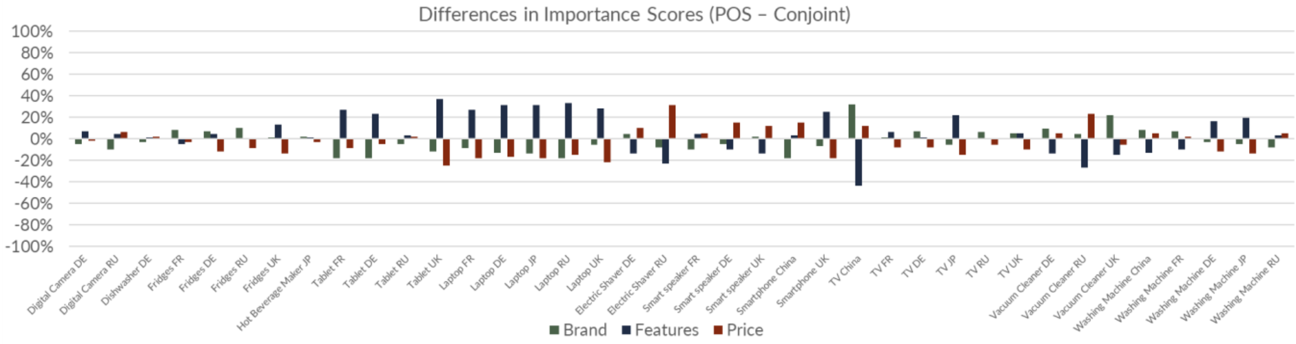


Figure 3: Attribute Importance: Hot Beverage Makers in Japan

	Conjoint	POS	Difference
Brand	15%	17%	2%
Feature	67%	68%	1%
Price	18%	15%	-3%

Figure 4: Attribute Importance: Tablets in UK

	Conjoint	POS	Difference
Brand	20%	7%	**-12%**
Feature	42%	80%	**37%**
Price	38%	13%	**-25%**

Figure 5 shows a summary of the differences in importance scores between the POS and conjoint, and how these compare to the results we would expect from random chance. 26% of the absolute differences in importance scores are 5% or less, 28% of differences are within 6-10% and 18% of differences are within 11-15%. Therefore, summing these, it means 72% of absolute differences between the conjoint and POS importance scores are within 15%, and 54% of differences are within 10%.

Figure 5: Summary of absolute differences in attribute importance (POS – Conjoint).

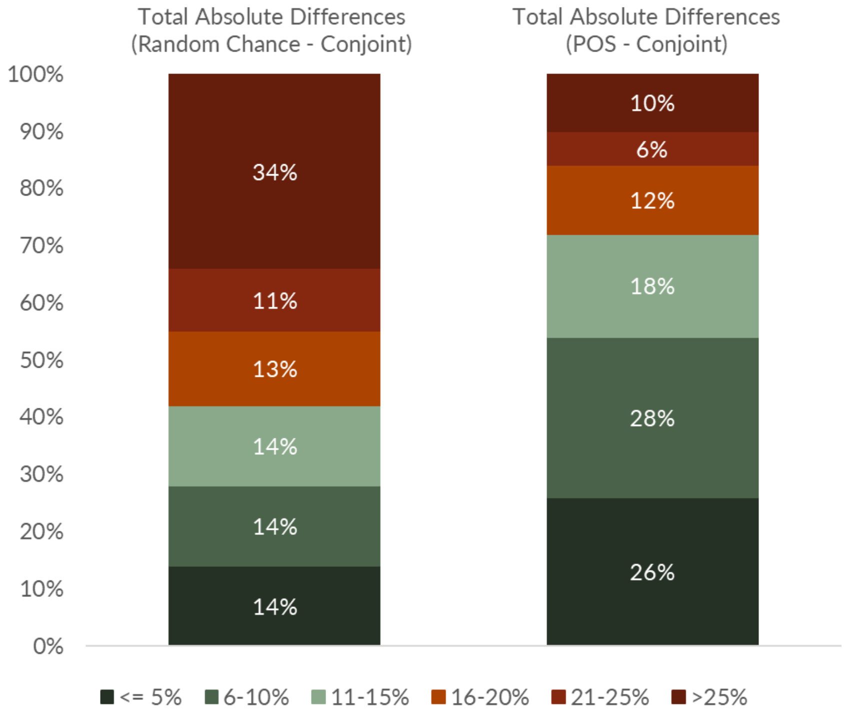
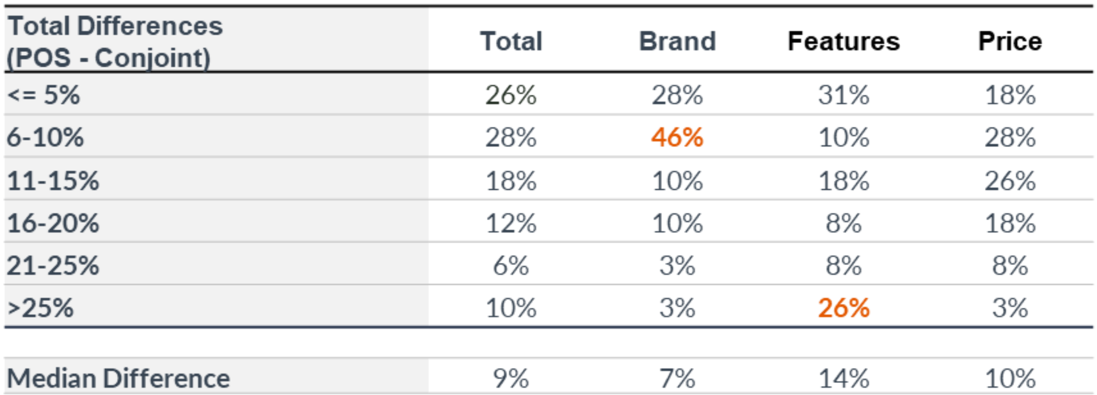


Figure 6 shows that differences in importance scores are smaller for brand but there are many large differences in feature importance, mainly due to the differences seen in Laptops and Tablets.

Figure 6: Absolute differences in attribute importance (POS – Conjoint) by attribute.



Total Differences (POS - Conjoint)	Total	Brand	Features	Price
<= 5%	26%	28%	31%	18%
6-10%	28%	46%	10%	28%
11-15%	18%	10%	18%	26%
16-20%	12%	10%	8%	18%
21-25%	6%	3%	8%	8%
>25%	10%	3%	26%	3%
Median Difference	9%	7%	14%	10%

Brand Preference

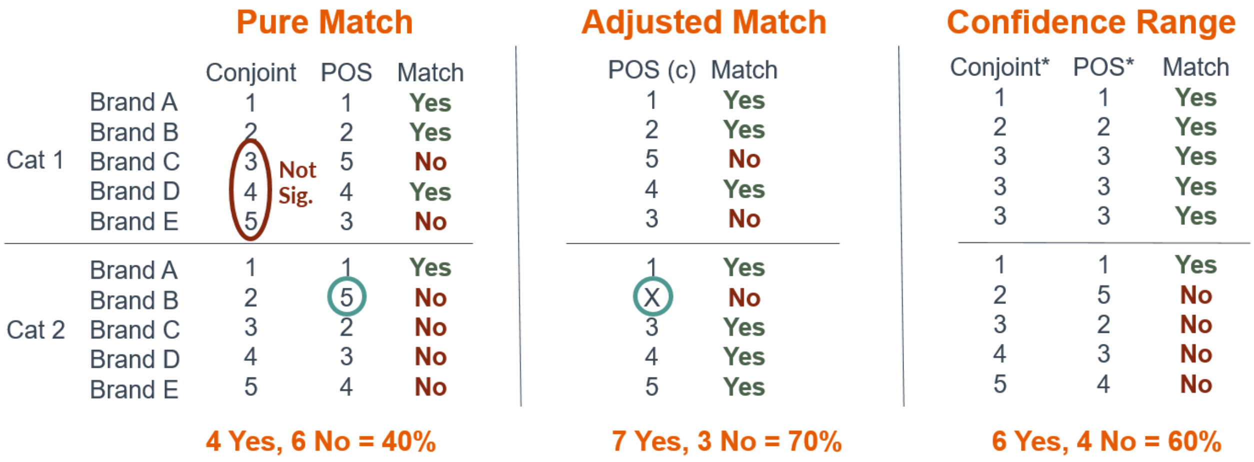
A simple and easy way to compare brand preference across the large number of categories is to compare the ranks in appeal of the brands from both the POS and conjoint models. Figure 7 shows the different ways we compare the ranks:

Pure Match simply calculates the proportion of ranks that match.

Adjusted Match accounts for the fact it only takes one of the rankings to be inconsistent between the POS and conjoint for all the other rankings to be out of line. We therefore make an adjustment where we ignore the rank that is out of line and renumber the ranks of the other brands.

Confidence Range takes into account the confidence interval of the conjoint utilities. We perform a t-test on the conjoint utilities and where we do not see significant differences between brand utilities, we assign the brands to have the same rank.

Figure 7: The different ways we compare ranks.



Pure Match

		Conjoint	POS	Match
Cat 1	Brand A	1	1	Yes
	Brand B	2	2	Yes
	Brand C	3	5	No
	Brand D	4	4	Yes
	Brand E	5	3	No
Cat 2	Brand A	1	1	Yes
	Brand B	2	5	No
	Brand C	3	2	No
	Brand D	4	3	No
	Brand E	5	4	No

4 Yes, 6 No = 40%

Adjusted Match

POS (c)	Match
1	Yes
2	Yes
5	No
4	Yes
3	No
1	Yes
X	No
3	Yes
4	Yes
5	Yes

7 Yes, 3 No = 70%

Confidence Range

Conjoint*	POS*	Match
1	1	Yes
2	2	Yes
3	3	Yes
3	3	Yes
3	3	Yes
1	1	Yes
2	5	No
3	2	No
4	3	No
5	4	No

6 Yes, 4 No = 60%

Figure 8 shows the results of matching the brand rankings. We achieve a pure match of 45%, over double what you would expect by random chance (20%). The adjusted match is higher at 51% and the confidence range match is higher still at 69%. Figure 9 shows the results of a qualitative assessment, where we manually looked at each category and subjectively appraised how well the ranks matched. 57% of categories were identical or very similar, 30% were "OK," and 14% showed large differences.

Figure 8: Comparison of brand rankings between POS and conjoint.

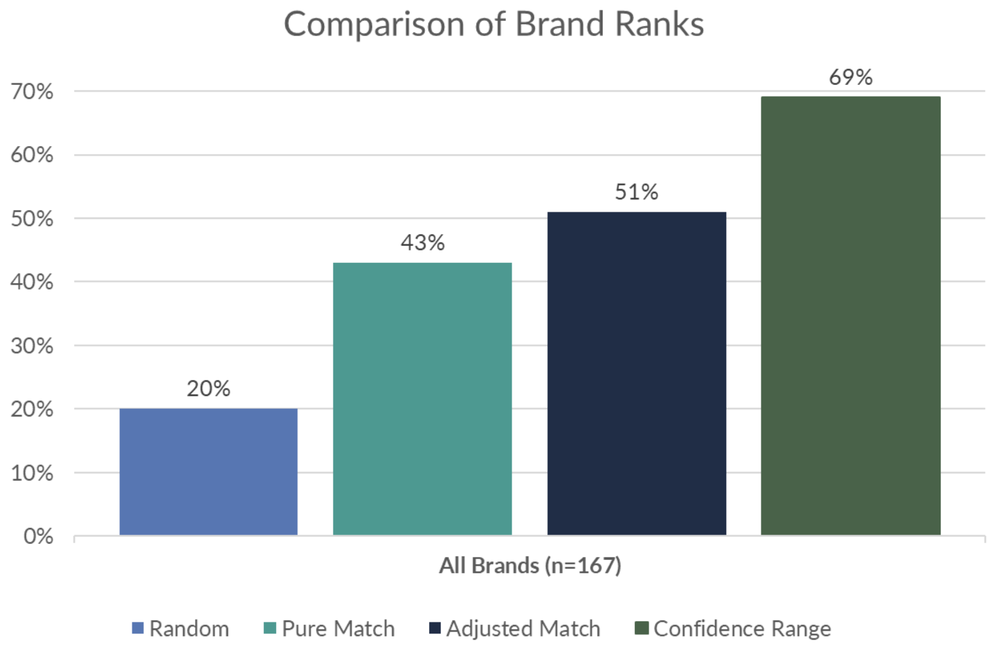
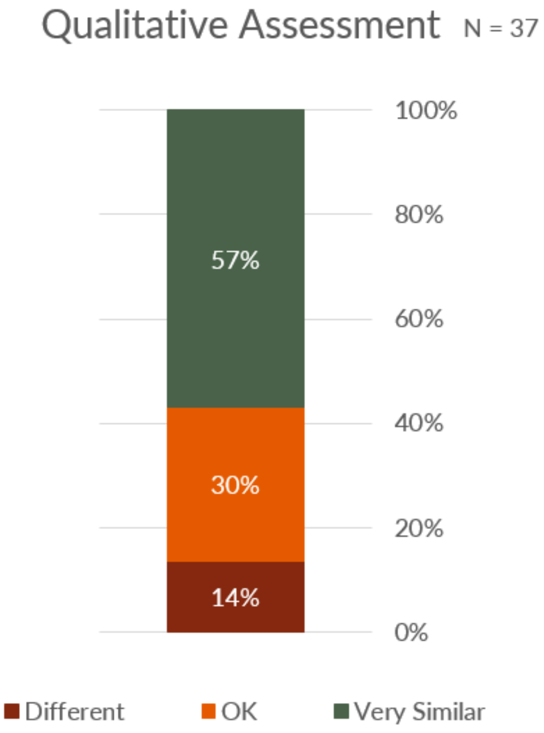
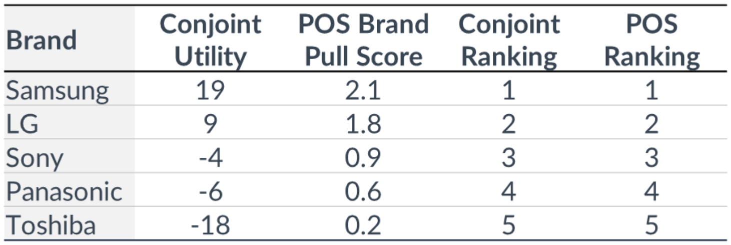


Figure 9: Qualitative comparison of brand rankings between POS and conjoint.



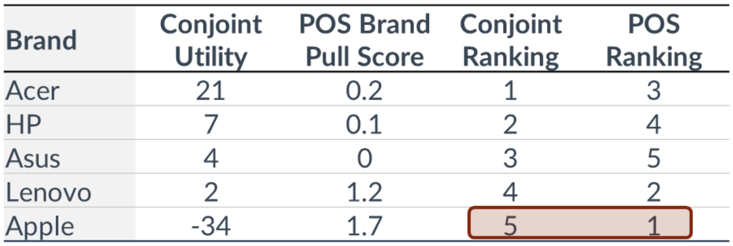
Figures 10 and 11 show results for a couple of individual cells. In TVs in the UK, the brand rankings are perfectly consistent. However, in Laptops in Germany, the brand rankings are inconsistent. In particular, we see a large difference for Apple where the POS model has Apple as the top ranked brand, and the conjoint has Apple as the worst brand. The appeal of Apple was much higher in the POS model compared to the conjoint across Laptops and Tablets in most countries.

Figure 10: TVs in UK



Brand	Conjoint Utility	POS Brand Pull Score	Conjoint Ranking	POS Ranking
Samsung	19	2.1	1	1
LG	9	1.8	2	2
Sony	-4	0.9	3	3
Panasonic	-6	0.6	4	4
Toshiba	-18	0.2	5	5

Figure 11: Laptops in Germany



Brand	Conjoint Utility	POS Brand Pull Score	Conjoint Ranking	POS Ranking
Acer	21	0.2	1	3
HP	7	0.1	2	4
Asus	4	0	3	5
Lenovo	2	1.2	4	2
Apple	-34	1.7	5	1

Feature Preference/Pricing

Figure 12 shows that, even though they measure different things, we achieve a pure a match of 69% between the conjoint feature preferences and the feature price premiums we calculate from the POS data.

Figure 12: Comparison of feature rankings between POS and conjoint.

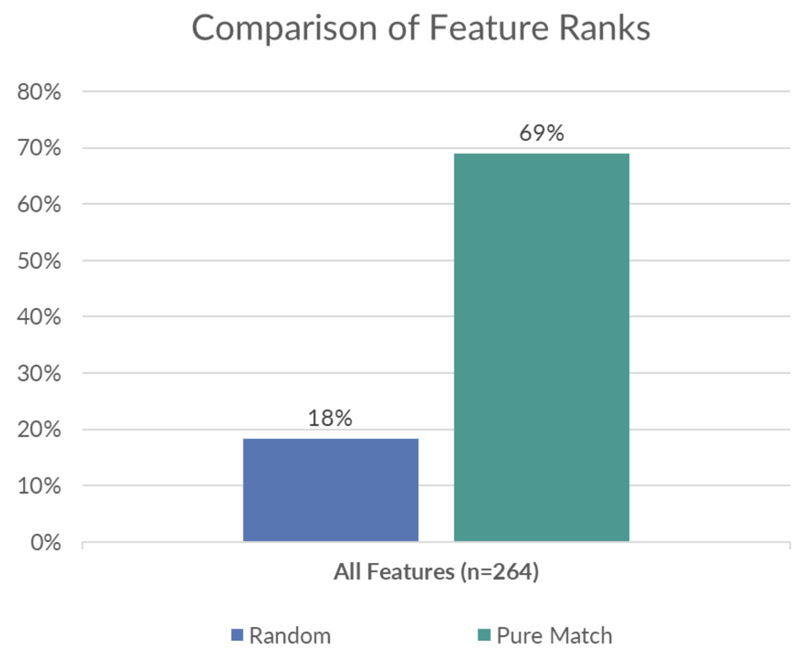


Figure 13 shows that 63% of the features are ordinal. By this we mean that the levels have a natural order of preference. For example, most consumers would prefer a camera with more megapixels than less. Figure 14 shows that we match these ordinal attributes at a much higher rate than the nominal attributes, which have no natural order of preference; 84% versus 43%, respectively.

Figure 13: Proportion of ordinal and nominal features in the analysis.

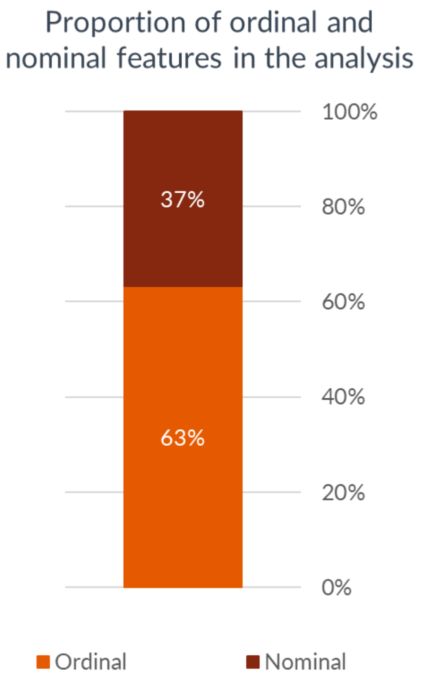
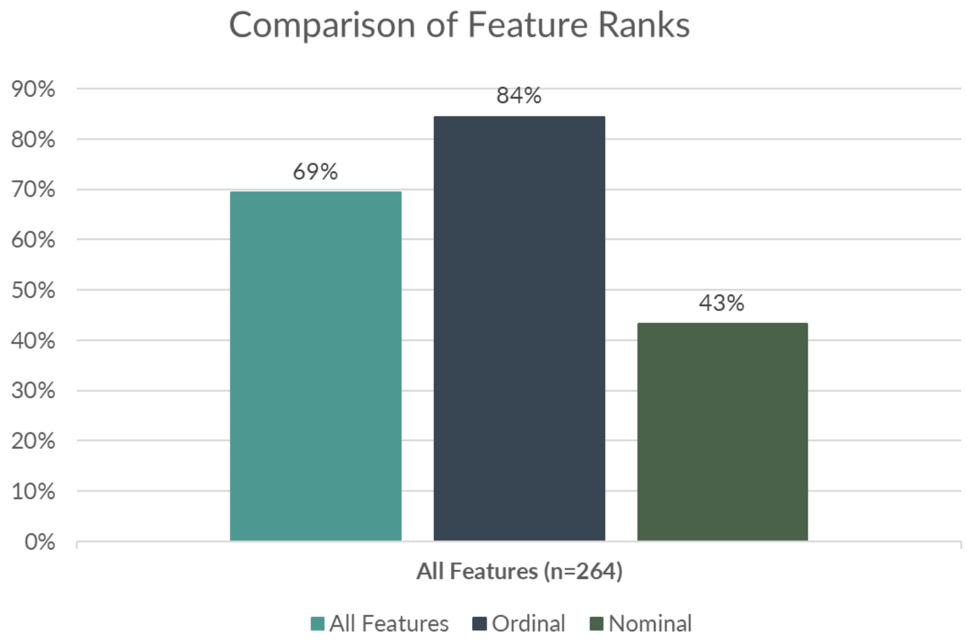


Figure 14: Comparison of feature rankings split out by ordinal and nominal features.



Price Elasticities

Figure 15 shows the conjoint produces much higher price elasticities in each category. Only in TVs in Germany is the category elasticity from the POS data (-1.2) comparable with the category elasticity from the conjoint (-1.5). Figure 16 shows the product price elasticities of all 38 products we ran models for do not correlate between the POS and conjoint models.

Figure 15: Comparison of category price elasticities between POS and conjoint.

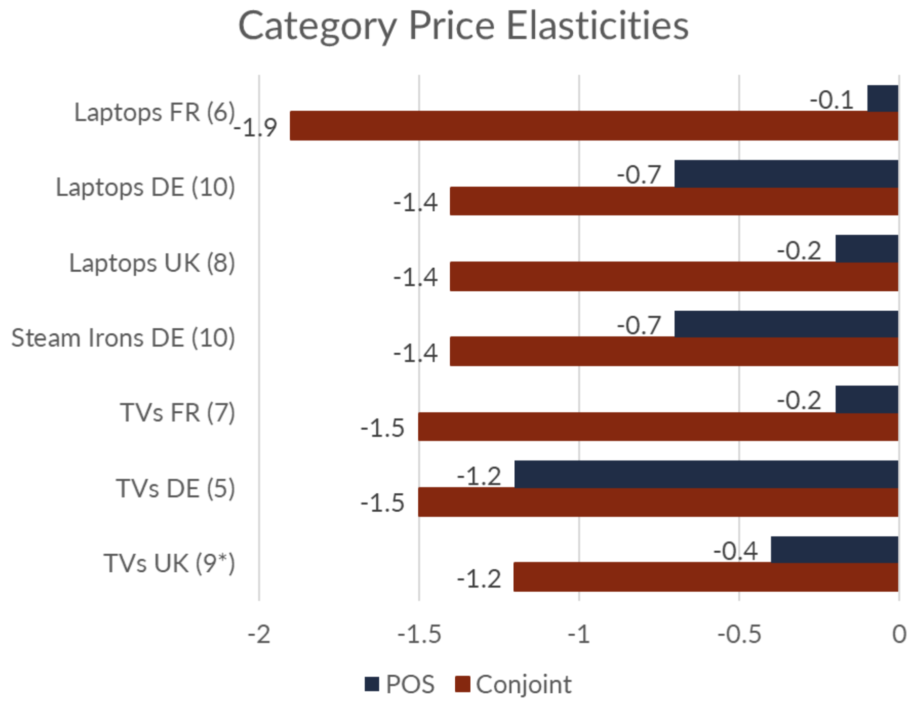
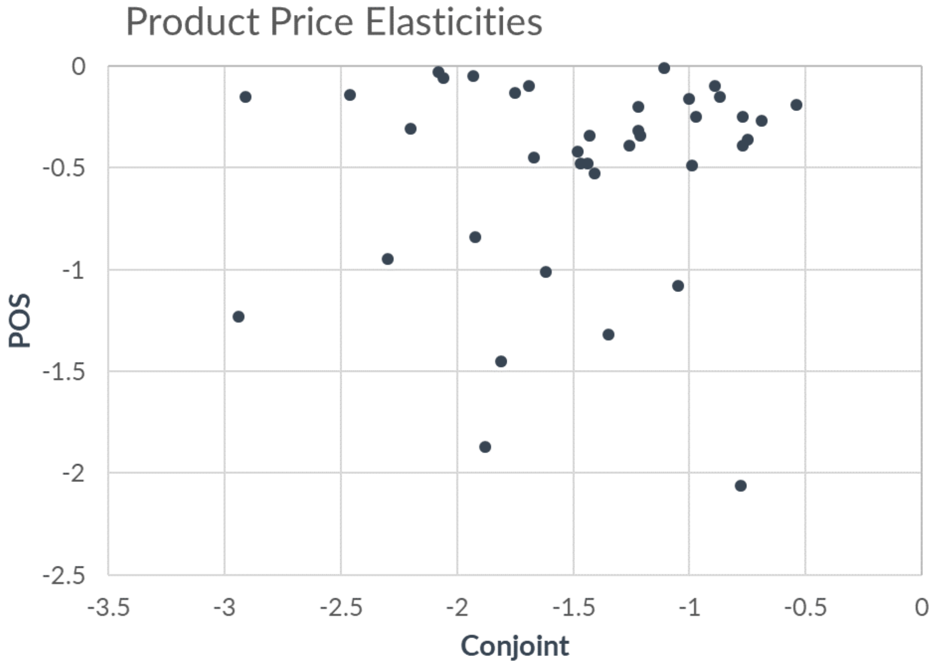


Figure 16: Correlation of product price elasticities between POS and conjoint.



Discussion

Attribute Importance

Attribute importance scores are reasonably consistent between the POS model and conjoint. It shows we are able to capture customer decision behaviour using POS data. The major differences are within Laptops and Tablets where feature importance is much higher in the POS models compared to the conjoint. This could be because of the limitations on the number of features we could test in the conjoint. In the conjoint exercise, we were limited to testing only six attributes relating to product features. However, in the POS model, we tested up to a maximum of 40 features (although this was lower for many of the categories). We therefore may miss off some features in the conjoint that are important in the purchase decision when buying a Laptop or Tablet. Hence, we underestimate the total feature importance in the conjoint for Laptops and Tablets.

Features account for around 60% of the purchase decision in both models, and features are ranked most important in all conjoint models and nearly all POS models. This shows that innovation in technology categories is crucial. Customers are willing to trade-off price and brand for a more technologically advanced product with the features they want. This is great news for market research, especially conjoint which can measure feature preferences, as it is important for manufacturers to know which features consumers look for in a product.

Brand Preference

Brand rankings are largely consistent between the POS data and conjoint. The major differences are again within Laptops and Tablets, where the appeal of the Apple brand is much higher in the POS models compared to the conjoint. There are several potential reasons for this. It is likely that there is a high in-store push and favourable in-store placement for Apple products. Such in-store activities are reflected in the POS data, leading to a higher preference for Apple, compared to the conjoint which does not take into account these external factors.

Another potential reason is that Apple only offers premium products which means taking the total brand sales across all Apple products in the POS model isn't effective at stripping out the influence of product features as is usually the case with other brands. We are not able to isolate the appeal of specifically the Apple brand from the appeal of the features contained in within the Apple products. We therefore may overestimate the appeal of the Apple brand in the POS model because part of the reason consumers choose Apple is because they like the features contained within the Apple product, not purely the Apple brand.

Apple is especially unique in form factor and design which may have much higher salience in-store, where a consumer is able to pick up and play with the product, compared to a conjoint setting where the respondent is not able to do this. Also, conjoint gives the respondent full market knowledge of the specifications and prices of all products and allows them to compare them on screen, side by side. However, in reality, many customers may not have such good market knowledge and will choose Apple because it is a brand they are more familiar with.

Finally, respondents can easily switch between brands in a conjoint setting. However, in reality, some customers will be locked within the Apple ecosystem as they own other devices with the same operating system. For example, if the consumer owns an iPhone, it may make them more likely to choose an Apple Laptop or Tablet so they can remain within the iOS ecosystem.

So, as we see in the case of Apple, the POS data is more suited to account for external factors, like in-store activities. When a consumer walks into a store they have a theoretical brand preference in their mind. This is what we measure in the conjoint. However, things can happen in-store that influence what brand they purchase. We see that brand importance is mostly consistent between the POS data and conjoint, so the importance of brand in the purchase decision is not greatly altered. However, the consumer may adjust their original theoretical preference for certain brands, as measured in conjoint, and switch to a different brand of similar preference due to in-store activities such as product placement, the behaviour of sales staff, and in-store (non-price) promotions like leaflets and banners.

This shows the importance of in-store merchandising and that manufacturers have enough room to influence the customer preference at the point of purchase.

Feature Preferences/Pricing

We are currently unable to measure consumer preferences for certain features using POS data because of the high multicollinearity between features. This is a big limitation currently, one which is not easily solved.

Interestingly, the rank orders of manufacturing prices and consumer preferences are consistent. This is largely due to a high proportion of features having a natural order of preference. It makes sense that manufacturers charge more for advanced features, so the prices from the POS data preserves the natural order of preference of the ordinal features. However, it does mean that manufacturers are doing a good job of pricing their products to match customer preference. This suggests they are using their expert knowledge of the category and perhaps, in some cases, they have conducted previous market research.

Conjoint remains the best way to measure consumer preferences for different features. Given that features are so important in the purchase decision within technology categories, this provides further evidence of the benefits of conjoint analysis to commercial clients.

Price Elasticities

We observe major differences in the price elasticities we generate from the POS models and the conjoint. Conjoint produces higher price elasticities than the POS models and price elasticities do not correlate between POS and conjoint. There are a number of potential reasons for this. Firstly, conjoint assumes all products are always available to consumers. However, we know from looking at the store-level POS data, in reality, only a subset of products are available in each individual store. There are therefore fewer alternatives for consumers to switch between when in-store. With fewer alternatives available, a consumer may become less sensitive to changes in price.

As previously mentioned, conjoint misses external factors such as in-store promotions, sales staff, and position on the shelf, which all affect the purchase decision a consumer makes in real life. These factors may cause a consumer to become less sensitive to price when in store. For example, you can imagine a scenario where a salesperson is pushing a product on a consumer, causing the consumer to care less about the price of the product.

The observed price variation in the POS data was very small for some products. Over the six-month modelling period, many of the products did not change their base price very much, less than +/- 5% in many cases. The lack of price variation makes it difficult to estimate price elasticity values. Restriction of price range and collinearity with other features could lead to lower estimates. For example, if a retailer only sold a small number of products at a given price, and they had similar features, the effect of price could be completely masked. There are also other external factors, such as in-store activities, we are not able to control for during the modelling process. Furthermore, in the conjoint, we tested a very wide range of prices because we wanted to cover the range of prices available across the whole market. However, this means the conjoint model is not sensitive to such small price changes, since we are simply linearly interpolating between two price points which are often far away from the price points being modelled.

We are not always able to accurately represent the products present in the POS data, in terms of the features they possess, with the levels tested in the conjoint. As we were limited by how many features we could test in the conjoint, we often misspecify the features of products. Furthermore, some features that may influence price elasticity, such as industrial design, physical look and feel, etc., may not have been tested in the conjoint at all. Therefore, the market scenarios we base the POS and conjoint models on are not always identical and, in some cases, they can be quite different. We may get more comparative results between the POS and conjoint if we used an SKU-Price conjoint approach, where each product in the POS model is modelled as its own separate level, allowing us to specify the exact features of the product to the respondent.

In the POS model, we explicitly strip out the effect of price promotions from the impact of changes in base price. This reduces the elasticity value we measure for the base price. Promotions were not tested in the conjoint, so we do not make the distinction between base prices and promotional discounts in the same way.

Finally, conjoint makes respondents fully aware of all prices of all products, and allows them to compare them on screen, side by side. However, in reality, many customers may not have such good market knowledge of prices. So, there is risk that the conjoint setting overinflates a consumer's price elasticity.

So, as we see in the case of brand preferences, the POS data is more suited to account for external factors, like in-store activities. When a consumer walks into a store they have a theoretical sensitivity to price in their head. This is what we measure in the conjoint. However, things can happen in-store that influence how sensitive they are to price.

This again shows the importance of in-store activities and that manufacturers have enough room to influence the customer preference at the point of purchase.

Benefits of POS Models

An obvious benefit of POS models is that there is no need to survey respondents. The ability to potentially save valuable time and money is becoming ever more pertinent as we increasingly live in a world where clients want insights delivered faster and at a reduced cost.

POS Models are based on real sales data. Rather than being based on what respondents say they will do in an artificial survey setting, they are based on what actually happened in the real world.

POS models take into account external effects which conjoint can't capture, such as in-store activities and distribution, both of which play an important role in a consumer's purchase decision. We see that a consumer's preferences and sensitivity to price can be altered depending on what happens when they are in-store. Although we can adjust for distribution at an aggregate level in a conjoint, we can do this at a much more granular level in the POS data as we know which products were sold in each individual store. This allows us to create models that are very close to the market reality, which is particularly useful when assessing how a product interacts with its competition. For example, when we calculate cross-elasticities, we only take into account the competitor products which were sold alongside the product of interest in each individual store.

Once trained, models can provide almost real time results on the latest market data. When using traditional surveys, the only way to get updated results is to interview a fresh sample of respondents which takes time and incurs extra cost.

Limitations of POS Models

POS data is not available to most market research agencies. However, there are plenty of other sources of sales data that could potentially be used to understand consumer preferences.

Although POS data is vast, some categories do not have enough purchases or have too much week-on-week variance to be used. We initially started with 68 cells but much of the modelling was based on 38 cells or fewer. We are often limited to only being able to model the top selling brands and products.

POS data only measures what you observed to happen in the market. This is particularly relevant when calculating price elasticities. The observed price variation was very small for most products. This limits the range of prices you can model, since you can only model with confidence the prices within the observed range. We also cannot model new brands, new products, or new product features like you are able to with conjoint.

The POS data is also backward-looking and is based on a specific set of market conditions. When making predictions about the future, we are assuming the same specific set of market conditions is present. However, this unlikely to be the case.

Due to high multicollinearity, it is currently not possible to measure feature preference using the POS data. This is big limitation and one that is not easily solved.

Due to the vast nature of POS data, significantly more computation power is needed to analyse the data compared to what is required when working with most survey data sets.

Conclusion

Conjoint and POS models differ in how they measure consumer preferences. Conjoint measures a theoretical preference, one that is not influenced by external market factors. However, the POS data takes into account the in-store realities and how these affect the purchase decision. POS models may therefore be beneficial when tactically modelling specific market scenarios as they are closer to market realities. However, POS models cannot be used for new product development, testing new features or new prices, or measuring feature preference. The best approach in such circumstances is still conjoint.

James Pitcher

Bradley Taylor

Dan Kelly

References

H2O.ai (2019) Variable Importance. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/variable-importance.html#variable-importance-calculation-gbm-drf

MAXIMIZING THE IMPACT OF OOH (OUTDOOR) ADVERTISEMENT USING DISCRETE CHOICE MODELING AND TEXT ANALYTICS

RAJAT GOEL[1]
RACHIN GUPTA[2]
STATWORLD RESEARCH SOLUTIONS

1. ABSTRACT

In this modern era of large numbers of players in every segment and tough competition in the markets, advertising plays an important role in the success of a brand. Out-of-home (OOH) advertising is considered one of the most important modes of advertising for various reasons. The changing lifestyle of consumers demands that these outdoor ads can capture their attention in just a single glance. This research uses Discrete Choice Modeling along with Text Analytics to create outdoor ads with maximum impact and likeability among the consumers. While Discrete Choice Modeling was used to pick the various visual elements of the advertisement, Text Analytics was used to create messages to be displayed in the advertisement, based on the sentiment of people for various words.

2. BACKGROUND

Out-of-home (OOH) advertising reaches the consumers while they are outside their homes. This type of advertising is focused on marketing to consumers when they are “on the go” in public places, in transit, or waiting in some commercial locations.

OOH advertising is an important mode of advertising for most organizations. It continues to outperform for various reasons; it is a mass reach medium, time spent outside the home is increasing, it remains unaffected by the erosion of audience due to proliferation of media channels, audience measurement is increasingly getting more sophisticated, it is physically present in the real world so it can’t be blocked like online ads, and many more.

As consumers are “on the go,” their attention span is very, very limited, sometimes only a few seconds. This demands an advertisement so carefully crafted that it can catch the attention of a maximum number of consumers in those few seconds, liked by the maximum number of consumers, in addition to creating the strongest impact on them with respect to the message the ad intends to deliver to consumers.

Typically, the pre-launch advertisement assessment is done using market research surveys. Respondents are shown advertisement concepts and are asked direct questions on various aspects such as overall liking for the advertisements, initial reaction, appeal of message statement, ratings of advertisement concept on uniqueness, relevance to them,

[1] Rajat is a Co-Founder and Director at StatWorld Research Solutions. He has an experience of 13+ years in marketing analytics, market research, logistics and supply chain. He specializes in choice modeling, pricing research, segmentation and other multivariate analyses. He has vast experience working with international clients on wide ranging market research and analytics assignments across multiple sectors.

[2] Rachin is a Co-Founder and Director at StatWorld Research Solutions. He has an experience of 16+ years in management consulting, marketing analytics, market research, and new business development. He has vast experience working with Fortune 500 clients on a multitude of consulting, analytics and market research assignments. He has worked in various strategic and entrepreneurial roles through his professional career.

effectiveness in terms of communicating the idea, likelihood to seek more information, etc. There are several limitations of this methodology as mentioned below:

- One can test only a limited number of advertising concepts, as respondents can evaluate only 2-3 advertisements with a fresh mind. Evaluating more than 2-3 advertisements brings respondent fatigue, leading to incorrect evaluation.
- Answering similar questions for advertisements with minimal differences doesn't bring out the real results from the consumer surveys, as respondents tend to give similar answers for all advertisement concepts.
- These are "stated responses" where respondents evaluate advertisement concepts one by one. Typically, the surveys don't provide them a platform to evaluate all the advertisement concepts simultaneously.

Discrete Choice Modeling combined with Text Analytics can not only improve the way advertisement assessment is done, but also addresses all the issues mentioned above. This paper explains how Discrete Choice Modeling can be used to increase the impact and likeability of these advertisements, and how Text Analytics can help create a better message, using an actual case study.

3. Study Design

3.1 Overview

This paper took inspiration from actual research done for a client from the tourism industry, operating in India. The client here is a travel services company who offers customized holiday packages for various locations across India. They are one of the major players in the travel services industry in India and enjoy the highest market share. Though they operate pan India, they are particularly strong in Southern India, especially in the state of KERALA, a state in the south of India.

Kerala has a long history of art and cultural heritage. Often called "God's Own Country," Kerala basks in the lap of nature. With a network of 44 rivers, the Arabian Sea in the west, and a channel of turquoise backwaters running throughout the state, Kerala is one of the most beautiful tourist destinations in India and is often thronged by tourists from all over the world.

The client wanted to tap consumers through some attractive outdoor advertisements, specifically promoting Kerala as a tourist destination.

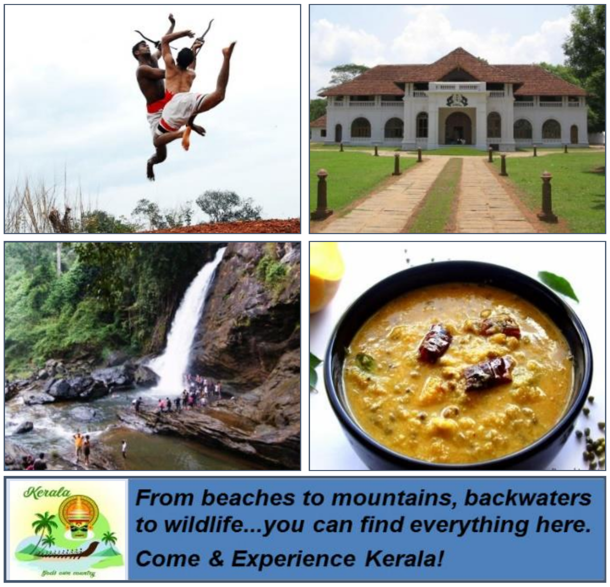
3.2 Initial Inputs

Kerala offers multiple entertainment options which include historical monuments, the scenic waterfalls, the great wildlife, beaches, its various art forms, the boat races, the cuisines, and many more. The client had a basic idea in mind about the advertisement and what all they wanted to have in it. These included images of any four entertainment options, a logo, and a message they intended to deliver to consumers. The advertisement concept the client had in mind is shown in Figure 1.

3.3 Limitations

Typically, the ad concepts are tested through surveys where respondents evaluate concepts on various aspects such as overall likeability, comprehension, message delivery, uniqueness, etc. However, there are limitations with this approach as one can evaluate only a limited number of concepts, as long surveys can cause respondent fatigue. Also, the ad concepts are created based on gut feeling of a few people and not on the basis of any scientific approach.

Figure 1: Advertisement Concept as per Client


Kerala
From beaches to mountains, backwaters to wildlife...you can find everything here.
Come & Experience Kerala!

In this particular situation, we were facing many challenges:

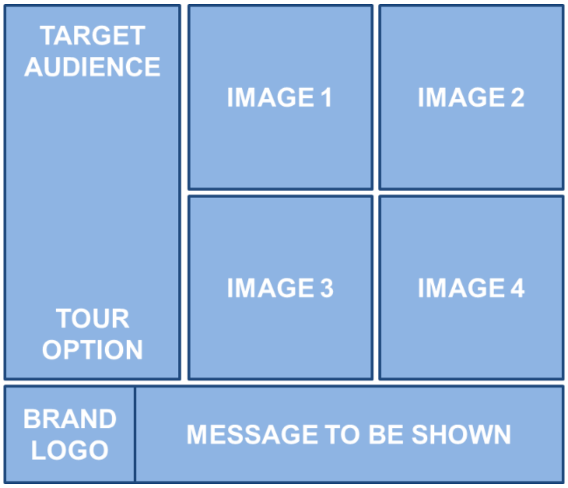
- Kerala offers as many as 10 different types of entertainment options of which we wanted to pick only four options.
- For each of the 10 entertainment options, we had 6-8 images. We wanted to pick one image for each of the four selected options.
- We wanted to evaluate multiple concepts and did not want to limit ourselves to evaluating only a few concepts.
- We also wanted to evaluate whether the positioning of images made any difference to the overall likeability of the advertisement and in the impact it creates.
- We wanted to create an impactful message innovatively rather than selecting any pre-decided message.
- We also wanted to address other points in the advertisement such as testing what is the most preferred duration of stay for tourists, and how we can address those issues to the relevant target audience through our advertisement.

3.4 Application of Discrete Choice Modeling

We wanted to address the limitations through an analytical approach. While figuring out which analytical technique could be best applied in this situation, we realized that our advertisement was actually a combination of multiple elements, i.e., the images that we needed to show in the advertisement, the audience we planned to target, the message that we intended to convey to consumers, the information that we wanted to display in the advertisement, and the brand display. These elements come together to make an effective advertisement. Based on this, we decided to apply Discrete Choice Modeling (Conjoint Analysis) to address the various questions.

We broke the client concept into its various elements and added a few more to create a blueprint for the final advertising concept. This has been shown in Figure 2.

Figure 2: Blueprint of Final Advertisement Concept with its Various Elements



We identified the attributes that could be a part of the conjoint experiment. This has been shown in Table 1.

Table 1: Attributes used for Conjoint

Broad Category	Attributes	Levels
Entertainment Options for Tourists	Art Forms	8
	Peaks	6
	Ayurveda	6
	Boat Races	7
	Cuisines	8
	Beaches	7
	Backwaters	6
	Monuments	7
	Waterfalls	8
	Wildlife	8
Target Audience	Target Audience	7
Tour Options	Tour Options	6
Positioning of Images	Positioning of Images	4

The levels of all attributes except "Tour Options" and "Positioning of Images" were in the form of images instead of textual values. At a given point, we were talking about managing about 70-80 images in the conjoint exercise. This brought in a lot of complexity. Also, there were 13 attributes that we were required to manage, which was quite a big number. To address these two challenges, we did the following:

- Due to the high number of "entertainment options" to be tested, respondents were asked to select their top 4 options before they answered the conjoint exercise.
- Only the respondent's selected 4 entertainment options were included in each conjoint exercise (in addition to the other fixed attributes).
- Further, "Positioning of Images" was programmed as a hidden variable.

This allowed us to have 7 attributes in the conjoint experiment, i.e., four selected "Entertainment Options," "Target Audience," "Tour Options," and "Positioning of Images." As "Positioning of Images" was a hidden attribute, respondents were effectively seeing 6 attributes in their conjoint exercise.

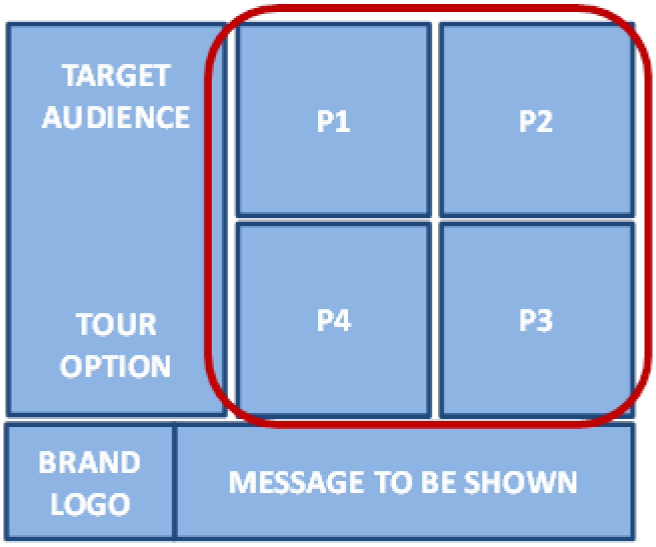
We decided to use the CBC technique. The other parameters of the conjoint exercise were:

- Partial-Profile Design
- 300 versions
- Complete Enumeration Method

The web-based survey was conducted among a population of 2,000 which included 500 from four different metro cities of India. The survey also probed respondents on their travel history, vacation preferences, and vacation frequency, etc.

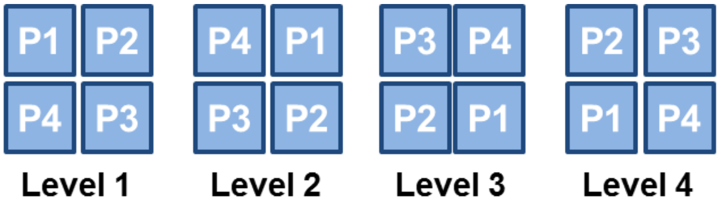
As mentioned earlier, we also wanted to evaluate if the position of images in the advertisement made any difference in the overall likeability and impact of advertisement. To address this, "Positioning of Images" was programmed as a hidden variable. P1, P2, P3, and P4 refers to the 4 image locations in the ad as shown in Figure 3.

Figure 3: Image Locations for "Positioning of Images" Attribute



In order to estimate the utility of any level of any of the ten attributes under “Entertainment Options,” at any of the four image locations in advertising concept, the levels of attribute “Positioning of Images” were defined as shown in Figure 4.

Figure 4: Levels for “Positioning of Images” Attribute



For level 1 of “Positioning of Images,” the four entertainment options pre-selected by the respondents were assigned P1 to P4, based on the order they were selected in. For exp., the option selected 1st was assigned P1, option selected 2nd was assigned P2, option selected 3rd was assigned P3, and option selected 4th was assigned P4. For the next level, P1/P2/P3/P4 were rotated in the clockwise movement.

Interaction effects between attributes under “Entertainment Options” and “Positioning of Images” were used to estimate the part-worth utility scores of any level of any attribute under “Entertainment Options” and any of the four image locations P1, P2, P3, and P4. After this, we had the part-worth utility scores of all levels of all attributes under “Entertainment Options” at all of the four image locations.

The logit report indicated that all standard errors were in the range of 0.02. This has been shown below.

```
Logit Efficiency Test Using Simulated Data
----------------------------------------------------------
Std Err Attribute Level
1 0.02573 1 1 Kalaripayattu
2 0.02571 1 2 Kathakali
3 0.02574 1 3 Koodiyattam
4 0.02577 1 4 Koothu
5 0.02568 1 5 Kutiyattam
6 0.02583 1 6 Mohiniyattam
7 0.02578 1 7 Pulikalli
8 0.02569 1 8 Theyyam
9 0.02135 2 1 Agastya Mala
10 0.02138 2 2 Anginda
11 0.02139 2 3 Banasura
12 0.02139 2 4 Brahmagiri
13 0.02133 2 5 Chembra
14 0.02137 2 6 Paithalmala
15 0.02138 3 1 Garshana
16 0.02141 3 2 Njavarkijhi
17 0.02135 3 3 Pizhichil
18 0.02133 3 4 Shirodhara
19 0.02139 3 5 Udvartana
20 0.02139 3 6 Abhyanga
21 0.02369 4 1 Aranmula
22 0.02370 4 2 Champakulam
23 0.02362 4 3 Indira Gandhi
24 0.02364 4 4 Kallada
25 0.02364 4 5 Kumarakom
26 0.02360 4 6 Nehru
27 0.02364 4 7 Payippad
28 0.02577 5 1 Appam
29 0.02582 5 2 Ela Sadya
30 0.02571 5 3 Fish Molee
31 0.02579 5 4 Idiyappam
32 0.02567 5 5 Parippu Curry
33 0.02576 5 6 Pumpkin Erissery
34 0.02571 5 7 Puttu
35 0.02589 5 8 Sadya
36 0.02373 6 1 Bekal
37 0.02361 6 2 Chowar
38 0.02365 6 3 Kovalam
39 0.02360 6 4 Marari
40 0.02361 6 5 Shankhumugham
41 0.02366 6 6 Varkala
42 0.02365 6 7 Vizhinjam
43 0.02132 7 1 Alleppey
```

44	0.02148	7	2	Kasargod
45	0.02129	7	3	Kollam
46	0.02146	7	4	Kumarakom
47	0.02137	7	5	Kuttanad
48	0.02129	7	6	Vaikom
49	0.02365	8	1	Anjuthengu Fort
50	0.02357	8	2	Bekal Fort
51	0.02372	8	3	Dutch Palace
52	0.02372	8	4	Jewish Synagogue
53	0.02363	8	5	Padmanabhapuram Palace
54	0.02364	8	6	Palakkad Fort
55	0.02363	8	7	Thalassery Fort
56	0.02579	9	1	Athirapally
57	0.02567	9	2	Chethalayam
58	0.02564	9	3	Keezharkuthu
59	0.02574	9	4	Meenmutty
60	0.02571	9	5	Palaruvi
61	0.02588	9	6	Power House
62	0.02587	9	7	Thommankuthu
63	0.02577	9	8	Vazhachal
64	0.02584	10	1	Aralam
65	0.02584	10	2	Choolannur
66	0.02592	10	3	Eravikulam
67	0.02570	10	4	Idukki
68	0.02562	10	5	Neyyar
69	0.02574	10	6	Peppara
70	0.02580	10	7	Periyar
71	0.02560	10	8	Begur
72	0.02356	11	1	Honeymoon
73	0.02372	11	2	Romantic
74	0.02371	11	3	Solo
75	0.02367	11	4	Biking
76	0.02355	11	5	Family
77	0.02361	11	6	Retirement
78	0.02377	11	7	Group
79	0.02137	12	1	4N/5D for INR. 9,999
80	0.02138	12	2	5N/6D for INR. 11,999
81	0.02133	12	3	6N/7D for INR. 13,999
82	0.02138	12	4	7N/8D for INR. 15,999
83	0.02135	12	5	8N/9D for INR. 18,999
84	0.02131	12	6	9N/10D for INR. 21,999
85	0.01591	13	1	1,2,3,4
86	0.01590	13	2	4,1,2,3
87	0.01593	13	3	3,4,1,2
88	0.01594	13	4	2,3,4,1
89	0.01969	NONE		

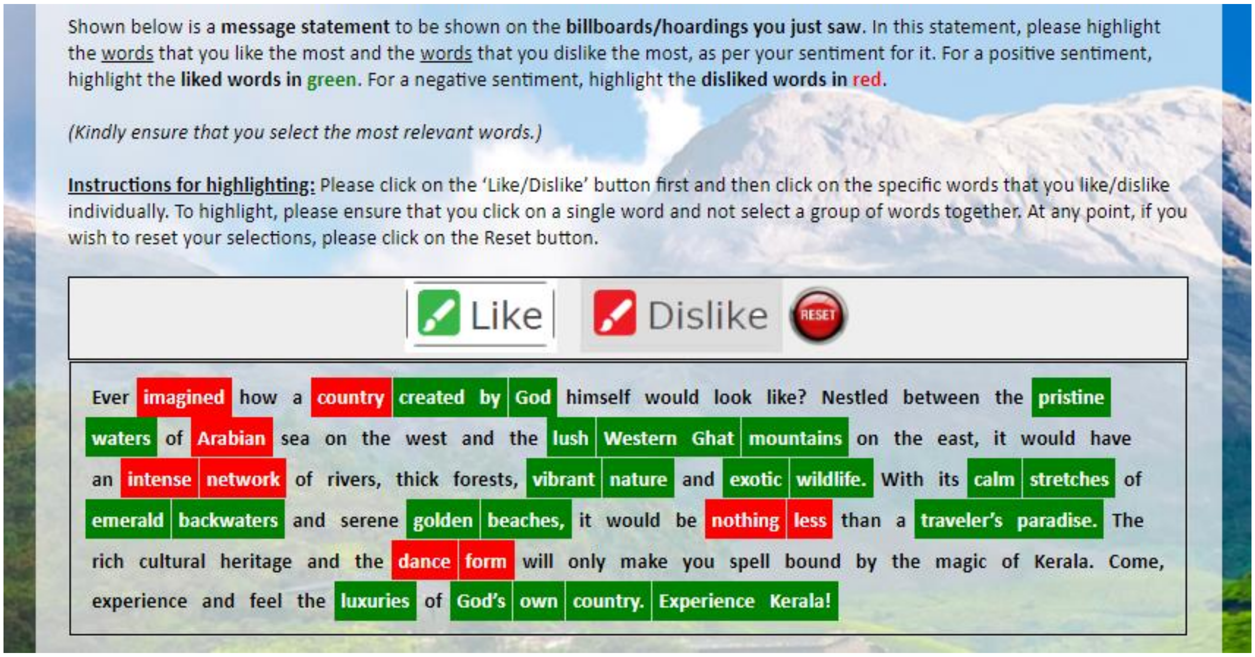
3.5 Application of Text Analytics

In a typical scenario, messages that we intend to show on the advertisement are developed by a team/individual and then they are put to the test by respondents, who evaluate those messages on various aspects such as overall impact, clarity, believability, comprehension, etc. We wanted to develop the messages somewhat innovatively rather than simply asking an individual/team to develop it based on their understanding.

We decided to use text analytics to create messages, wherein we planned to pick words on the basis of consumer sentiment for those words. We showed them a message and asked them to pick words from the message for which they had a positive sentiment. Similarly, they picked the words for which they had a negative sentiment. This was done using a "Text Highlighter" exercise, where the respondents were able to select the words they liked in green and words they disliked in red. This is shown in Figure 5.

Based on the sentiment mentioned for various words by respondents, the responses were aggregated. The cumulative responses for all participants allowed us to create frequency reports for each and every word in the tested message, based on which we estimated the positive sentiment as well as the negative sentiment for any word, thereby calculating the net sentiment of all words. At the end of this exercise, we had the net sentiment of all words and relative intensity of sentiment of all words. Using this information, we picked words with the strongest positive sentiment to include in the message. The new message was then put to test to respondents for evaluation on overall impact, clarity, believability, comprehension, etc.

Figure 5: Text Highlighter Exercise



Shown below is a **message statement** to be shown on the **billboards/hoardings you just saw**. In this statement, please highlight the words that you like the most and the words that you dislike the most, as per your sentiment for it. For a positive sentiment, highlight the **liked words in green**. For a negative sentiment, highlight the **disliked words in red**.

(Kindly ensure that you select the most relevant words.)

Instructions for highlighting: Please click on the 'Like/Dislike' button first and then click on the specific words that you like/dislike individually. To highlight, please ensure that you click on a single word and not select a group of words together. At any point, if you wish to reset your selections, please click on the Reset button.

Ever imagined how a country created by God himself would look like? Nestled between the pristine waters of Arabian sea on the west and the lush Western Ghat mountains on the east, it would have an intense network of rivers, thick forests, vibrant nature and exotic wildlife. With its calm stretches of emerald backwaters and serene golden beaches, it would be nothing less than a traveler's paradise. The rich cultural heritage and the dance form will only make you spell bound by the magic of Kerala. Come, experience and feel the luxuries of God's own country. Experience Kerala!

4. Results and Discussion

This research shows that analytics finds an application in creating advertisements that can have relatively higher likeability and impact among consumers. The elements that came together to make the advertisement were selected based on Discrete Choice Modeling in such a way that the "Utility" of advertisement was maximized to the consumers. The message to be displayed on the advertisement was crafted based on the sentiment of people for various words, before being evaluated by people on various aspects.

4.1 Discrete Choice Modeling Results

As mentioned above, the advertisement in this case was made up of various elements. The idea of doing conjoint analysis in this case was not to compare a few advertising concepts and to calculate some sort of preference share for those concepts. The idea of doing conjoint analysis here was that we wanted to pick those elements in the advertisement which together would maximize the "Utility" (or desirability) of the advertisement to consumers.

Conjoint analysis helped us in picking 4 entertainment options out of 10. We further picked an image each for these four entertainment options from 6-8 images available for each of them, on the basis of conjoint analysis.

The images for target audiences tested in Conjoint were of different themes. For example, we had an image showing an elderly couple, an image showing a young couple, an image with a group of young friends, an image with a family on vacation, etc. We wanted to test whether showing the corresponding image to the corresponding type of person makes more sense or if we should have a generic image, i.e., will the image of romantic couple resonate more with a romantic couple and similarly for others?, or will a generic image such as one for a family be equally impactful? Conjoint analysis helped us in evaluating this and we were able to pick an image that aptly addressed the target audiences with high impact.

Conjoint analysis also helped us put the most preferred tour duration option on our advertisement, i.e., should the ad display a tour of 5 nights at XX$ or a tour of 8 nights at YY$ or a tour of 10 nights at ZZ$.

One of the most important aspects that we were able to evaluate through conjoint analysis was whether the position of four entertainment options in the advertisement made any difference to the overall impact and likeability of the advertisement; for example, whether the image for any selected entertainment option created more utility to consumers if it was at P1 or if it was at P2/P3/P4? This allowed us to identify image locations for the four entertainment options that were selected on the basis of conjoint analysis.

4.2 Text Analytics Results

The text analytics allowed us to create very refined and concise messages which were based on the true positive sentiment of people. The words that formed the message were liked by people, and hence we believed that the multi-word messages would also be liked. As a part of a separate exercise, these messages were evaluated on multiple aspects by consumers before we decided the final message that was to be shown on the advertisement. The original message that was shown to respondents for selecting their sentiment for various words and what we arrived at finally are shown below in Figures 6 and 7, respectively.

Figure 6: Original Message Evaluated in the Survey

You don't need to plan separate vacations to enjoy lofty mountain ranges or a beach destination or a wildlife safari. Come visit Kerala and explore the wonders that await you here. From golden beaches and high mountains to emerald backwaters and powerful art forms, you get many choices to create and take home memories. Experience Kerala!

Figure 7: Message Created Based on Text Analytics

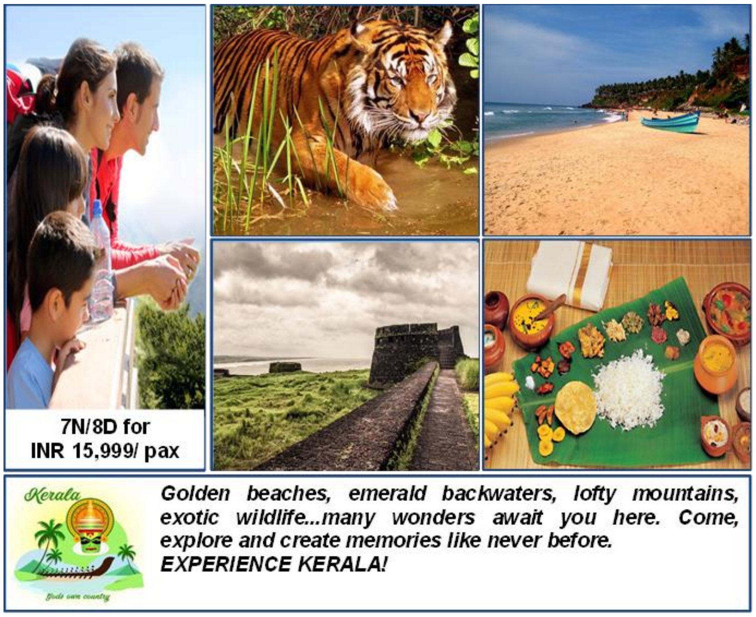
Golden beaches, emerald backwaters, lofty mountains, exotic wildlife...many wonders await you here. Come, explore and create memories like never before.

Experience Kerala!

5. Conclusion

The research allowed us to create an advertisement which not only addressed all the limitations we had at hand, but also had the potential to have maximum likeability and impact among consumers. What we had originally (Figure 1) and what we arrived at finally (Figure 8) were strikingly different, and the final product looked much better and more impressive.

Figure 8: Final Advertisement Created Based on Analysis



7N/8D for
INR 15,999/ pax

Golden beaches, emerald backwaters, lofty mountains, exotic wildlife...many wonders await you here. Come, explore and create memories like never before.
EXPERIENCE KERALA!

The final advertisement had the following features:

- The highest combined utility of advertisement to consumers.
- Compared all combinations of attributes to arrive at the best advertisement.
- Impactful message based on viewer sentiment.
- Addressed most relevant target audience.
- Provided required information to consumers in a comprehensive manner.

6. Recommendations

The key question that is often asked is whether this or similar analysis can be applied in all cases, i.e., if similar analysis can help improve any other advertisement. Our answer to this is "Yes." It can be applied to any such case where we can clearly identify and separate the elements that make up the advertisement. If we are able to do that, this analysis can very well be adapted to improve that advertisement.

Rajat Goel

Rachin Gupta

USING ADAPTIVE CHOICE-BASED CONJOINT ANALYSIS TO UNRAVEL THE DETERMINANTS OF VOTER CHOICES

DAVID BAKKEN
FORESEEABLE FUTURES GROUP
GRETCHEN HELMKE
UNIVERSITY OF ROCHESTER
MITCH SANDERS
MELIORA RESEARCH

INTRODUCTION

Since the 1950s, with the research that led to publication of *The American Voter* (Campbell, Converse, Miller, and Stokes, 1960), political scientists have conducted systematic, survey-based research to understand voting behavior. While there are different theoretical frameworks for understanding voter behavior (sociological approaches, psychological approaches, and economic approaches), there is general agreement that candidates are defined by multiple attributes, including party alignment and positions on key issues; demographics like gender and ethnicity; personal and political experiences; and perceived characteristics like competence, integrity, morality, and compassion.

Voters' preferences are typically construed as a function of party identification, general ideological orientation, beliefs or feelings about specific policy issues, and their overall perceptions of individual candidates. For example, Bartels (2018) posits two attitudinal dimensions, *limited government* and *cultural conservatism*, to explain the main differences between and within the two major political parties in the US. These orientations undoubtedly shape voters' party and policy preferences.

In tandem with our exploration of policy preferences, and building on the work of Graham and Svolik (2019) and Svolik (2018), we investigate the degree to which voters might trade off commitment to democratic principles (e.g., electoral fairness, checks and balances, freedom of the press) in favor of voting for one's preferred party and ideologically consistent policies.

To see how these trade-offs might operate, consider a primary situation where a strong Republican voter with high scores on both the limited government and cultural conservatism dimensions is asked to choose between a candidate who supports moderate policies on immigration and the environment, conservative tax and health care policies, and endorses key democratic principles (e.g., "Elected officials must obey the courts even when they think the decisions are wrong") and a candidate with very conservative positions on all policies but less support for democratic principles (e.g., "Elected officials should not be bound by court decisions they regard as political"). Will this voter choose the more ideologically consistent candidate with "undemocratic" principles or select the less ideologically consistent candidate who upholds the democratic principles?

In our study, we use Adaptive Choice-Based Conjoint analysis to understand the impact of a candidate's party, specific policy positions, and orientations toward democratic principles on individual choices between candidates. As part of our analysis, we derive individual-level preferences for general democratic principles and for specific policy positions, in order to understand potential tradeoffs between them.

Conjoint Analysis in Political Science

Conjoint analysis has only recently become popular in political science. Hainmueller, Hopkins and Yamamoto (2014) are generally credited with introducing conjoint analysis to political science. The primary appeal of conjoint analysis to these researchers lies in the ability to make causal inferences through random assignment. Political scientists have relied on such survey experiments to understand the effects of variables of interest on outcomes such as voter choices.

In the classical survey experiment, respondents are randomly assigned to different versions of the survey, with the objective of comparing the effect of different "treatments" presented in the different survey versions. In theory, random assignment to treatments makes it possible to rule out the effect of unobserved or omitted variables, including selection bias and other endogenous factors.[1] However, in many experiments it is difficult or impossible to isolate the individual contributions of the different components of a *multidimensional* treatment.[2] Conjoint analysis offers a way to decompose the overall effect into the individual contributions of the separate components of a treatment.

Hainmueller et al. report on two separate conjoint experiments. In the first experiment, they varied eight objective characteristics of would-be (hypothetical) presidential candidates. Six of these attributes had six levels each, while the remaining two had two levels each. For example, the candidate's profession had the following levels: business owner, lawyer, doctor, high school teacher, farmer, car dealer.[3] Military service, in comparison, had two levels: "served" and "did not serve." Survey respondents were presented with six different pairs of candidates. In addition to choosing between the candidates, respondents rated each candidate on a 7-point scale that indicated the strength of their support for each candidate.

In the second experiment, a different sample evaluated profiles of prospective immigrants to the United States. Key attributes included country of origin, reason for applying for entry, prior trips to the U.S., language skills, profession, and education level. Again, respondents saw two profiles at a time, made a choice between the profiles, and then rated each profile on a 7-point scale to indicate how strongly they felt each immigrant should or should not be granted entry. Whereas the candidate experiment was a completely orthogonal design, for the immigrant experiment prohibitions were included such that certain professions required minimum levels of education and certain reasons for applying (e.g., "fleeing persecution") were restricted to countries of origin where those reasons were at least plausible.

[1] In practice, selection- and treatment-related attrition can transform a randomized experiment into a quasi-experiment, where other techniques must be employed to account for such effects.

[2] To illustrate this, Hainmueller et al. cite an experiment conducted by Brader, Valentino and Suhay (2008) in which the researchers randomly varied two aspects of an otherwise identical news article. One aspect was the ethnicity of an immigrant described in the article. Ethnicity was operationalized along three dimensions: country of origin, face, and name, to create two distinct ethnic identities (one Russian and one Mexican). Given the design, it is impossible to determine the relative contribution of the three dimensions that defined ethnicity.

[3] While these levels encompass a wide range of professions, only one, "business owner" (without additional political experience), appears to be represented among actual presidential candidates.

Hainmueller, Hangartner, and Yamamoto (2015) conducted a subsequent study in which they compared conjoint survey experiments and vignette experiments to real-world behavioral benchmarks using a variation of the immigrant experiment described above. They found that the paired conjoint design where two immigrant profiles are compared in each task performed well in recovering the effects of the attributes as observed in the behavioral benchmark.

More recently, Bright Line Watch, a group of political scientists at the University of Chicago, Dartmouth College, and the University of Rochester, conducted a conjoint study to explore the trade-offs that voters make between upholding democratic values versus partisan and policy preferences (Carey et al., 2019). They included candidate attributes (gender, ethnicity, and party affiliation), policy positions or beliefs on taxation and discrimination, and candidates' support for selected democratic principles (bipartisan cooperation, voter access, independent investigations into misconduct by elected officials, and independence of the judiciary). They found that while partisanship had a large marginal effect on candidate choice, respondents from both political parties penalized candidates who supported the position that elected officials should not have to obey court decisions that were politicized. Republican respondents penalized candidates with progressive tax policies, while Democrat respondents rewarded those candidates.

In all of these studies, the effect of interest is the "average marginal component effect" (AMCE) which is the difference in the probability of choosing a profile (in the discrete choice case) with one value of a selected attribute versus profiles with different values of that attribute, averaged across all values of the other attributes in the model. This would be analogous to a main effect estimated using analysis of variance for a multivariate factorial experiment, or the β coefficient estimated via regression with dummy variable coding for the attributes. The AMCE is estimated at the population level. The estimation of AMCE is derived from the framework of potential outcomes and is non-parametric.

While showing promise for the use of conjoint analysis in political science, these studies have some potential limitations that might be addressed through different conjoint designs or estimation approaches. For example, with paired conjoint designs (two alternatives per task) the overall design is likely to be sparse. In the Hainmueller et al. (2014) candidate experiment, there are 186,624 possible candidate profiles but each respondent saw only 12 different profiles.

Second, aggregate estimation limits the extent to which preferences of different groups of respondents can be investigated. In the Bright Line Watch study, for example, separate aggregate models were estimated for each of the four subgroups of interest. Similarly, with aggregate estimation we cannot see the multivariate distribution of preferences in the way that disaggregate estimation (e.g., hierarchical Bayes or finite mixture models) enables.

Rather than avoid making assumptions about the underlying process that generated the data, we might want to test different hypotheses about that process. Hainmueller et al. specifically reject the additive compensatory model that is widely used in conjoint analysis for marketing applications. However, rather than ignore the underlying decision process, we may want to test or otherwise account for the presence of screening heuristics, for example.

In the presence of screening rules or other heuristics, the pairwise designs of these studies could fail to capture the marginal effects of some attributes. For example, to the extent that partisanship is a dominant or "must have" attribute for at least some respondents, information about the true independent preferences for candidate characteristics or policy positions may not be revealed in the choices made.

Using ACBC to Measure Voter Preferences

Some relatively simple modifications to the CBC designs and estimation methods that have been used by political scientists can resolve or mitigate some of the limitations described above. Expanding the number of choices to three or four candidates with balanced overlap designs will reduce the effects of a "must-have" attribute level. (This method insures that respondents will occasionally have to choose between options that each have the same "must have" feature.) Disaggregate estimation using HB (with or without covariates) or latent class, for example, would provide insights into the distribution of policy and candidate preferences across the target population.

Based on earlier findings that partisanship is a strong determinant of candidate preferences, we chose to use Adaptive Choice-Based Conjoint analysis to understand preferences for policy positions. We hypothesized that voters may apply strong screening rules in their choice of candidates. Republican voters, for example, might reject all Democratic candidates and all progressive tax policies. Similarly, Democrats might reject all Republican candidates and all proposals to restrict voting access.

In a standard CBC experiment, respondents are presented with choices that are largely determined by balanced and independent combinations of attributes to create unique profiles that are not influenced by prior information obtained from respondents. In contrast to this, Adaptive Choice-Based Conjoint gathers information from respondents which is used to shape the choices that are presented to respondents.

The first step in this process is usually a "build your own" (BYO) exercise. Respondents pick their preferred level of each attribute in order to configure their ideal product or service—or perhaps political candidate. The next step presents respondents with profiles that are constructed by an algorithm that seeks near-orthogonality in the combinations of attributes, with the levels selected in the BYO exercise oversampled, while other attribute levels are sampled equally. In this step (the screening section), respondents indicate, for each profile, whether it is a possibility or not. The researcher can insert *must-have* and *unacceptable* questions. Attribute levels that are always present in concepts that respondents state are a possibility will appear in must-have questions. Levels that are never included in acceptable profiles will appear in the unacceptable questions. Any attribute level identified as unacceptable will be excluded from subsequent profiles.

Profiles that are deemed acceptable in the screening section are carried into the tournament section, which resembles a typical single choice CBC experiment. Researchers can add a final set of calibration questions where respondents indicate their likelihood of choosing selected profiles if they were to be available in the market.

ACBC was appealing because we hypothesized that party affiliation may be a screening attribute for candidate selection. Furthermore, the BYO exercise would provide a direct measure of the appeal of the different policy positions we set out to test. Because we are specifically interested in the degree to which voters might trade off their preferred policy positions and democratic values when choosing a candidate, the ability to compare their BYO picks with their choices in the tournament exercise appealed to us.

METHODOLOGY

We conducted an online survey with a general population sample of 1,005 US adults (age 18 and up). The sample was drawn to reflect the population distribution of US adults with respect to sex, age, and education. Following the example of the Bright Line Watch study, we did not restrict the sample based on voter eligibility or registration. We fielded the survey from July 1 to July 3, 2019 (approximately one week after the first televised debates among contenders for the Democratic party presidential nomination).

In addition to the ACBC exercise we included several questions about party affiliation, support for President Donald J. Trump, ideological orientation, and participation in various activities that indicate political engagement.

For the ACBC attributes we selected a set of six issues that are prominent in the current political discourse: healthcare, immigration, taxation, the environment, the social safety net, and voter access. We also selected four democratic values (out of the larger set that Bright Line Watch measures on a regular basis). The complete set of attributes and levels is shown in Table 1.

Table 1

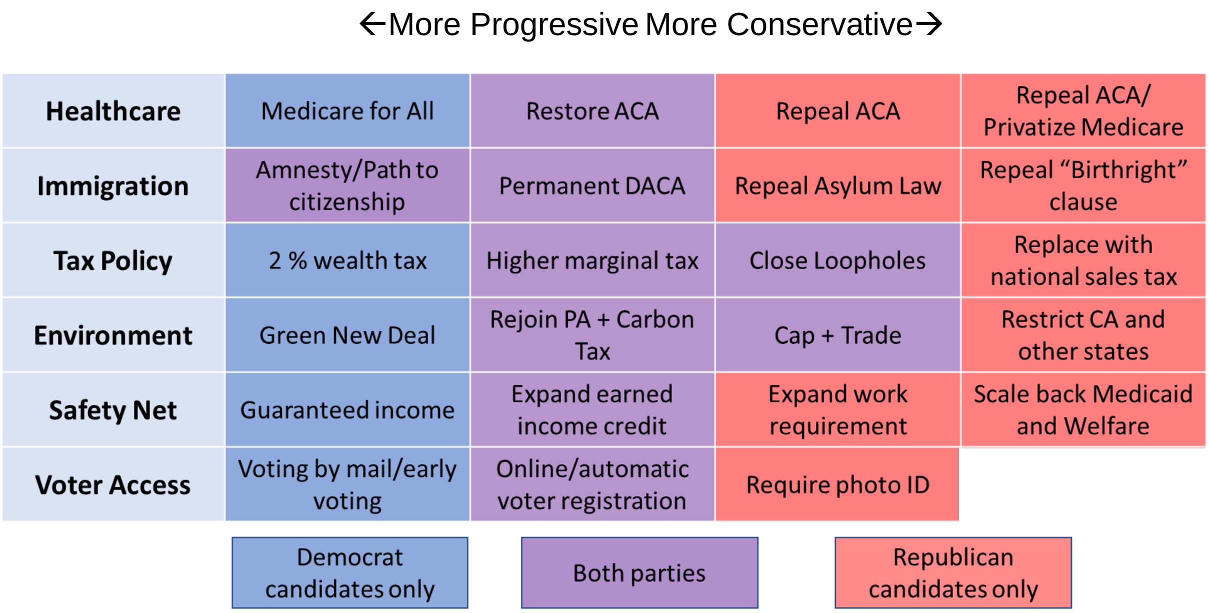
Attribute	Attribute Levels
Healthcare	• Medicare for all • Restore Affordable Care Act and add public option • Repeal Affordable Care Act • Repeal Affordable Care Act and privatize Medicare
Immigration	• Amnesty with a path to citizenship for undocumented individuals who can show five years of economic contribution • Make Deferred Action for Childhood Arrivals (DACA) permanent and increase the number of immigration judges to process asylum requests • Repeal the law that allows immigrants who reach US by any means to request asylum • Repeal the “birthright citizenship” clause of the Fourteenth Amendment to the US Constitution
Tax Policy	• 2% tax on the wealth of the 75,000 richest families in America • Raise marginal tax rate on income above $400,000; restore inheritance taxes to Clinton era levels • Close loopholes used by wealthy taxpayers and give the IRS more resources to enforce the rules • Replace income tax with a national sales tax
Environment	• Green New Deal • Rejoin Paris Accord, restore Obama-era regulations, implement Carbon Tax • Implement “cap and trade” for carbon emissions

	• Prohibit California and other states from enacting regulations that are stricter than US
Social Safety Net	• Guaranteed minimum income • Expand earned income tax credits to cover more people • Expand work requirement as a condition for receiving benefits • Scale back Medicaid and Welfare programs
Voter Access	• For national elections, require states to permit voting by mail and early voting • Permit states to offer online voter registration and automatic registration when applying for a driver's license • Require voters to present a photo ID when voting in national elections
Bipartisan cooperation	• Promise to seek bipartisan compromise • Promise to stick tightly to party's principles
Independence of the press	• The government should not interfere with journalists or news organizations • The mainstream media are irresponsible and should be constrained
Gerrymandering	• Drawing legislative districts to give one party an advantage should be prohibited • The winning party should have the right to set the boundaries of legislative districts
Judicial deference	• Elected officials must obey the courts even when they think the decisions are wrong • Elected officials should not be bound by court decisions they regard as politicized
Candidate's Party	• Republican • Democrat

One fact of partisan politics is that the parties espouse different policy positions. We sought positions for each issue that represent an extreme left or progressive stance, a moderate left or progressive stance, a moderate right or conservative stance, and an extreme right or conservative stance. We also looked for positions that had been expressed in party platforms, by individual candidates, or by left- and right-leaning think tanks. The policy statements for the four democratic values (bipartisan cooperation, independence of the press, gerrymandering, and judicial deference) were adapted from the Bright Line Watch survey.

In the candidate conjoint experiment conducted by Bright Line Watch there were no conditional relationships between policy statements and party affiliation. Thus, both Republican and Democrat candidates could endorse, for example, either a more progressive tax policy or a less progressive tax policy. In the real world it is extremely unlikely that a Republican party candidate would support a very progressive policy, and equally unlikely that a Democrat candidate would support a very conservative policy. We took advantage of the alternative-specific design capabilities of ACBC to restrict the most extreme policy positions to the appropriate party. Figure 1 shows the mapping of policy positions to candidate's party affiliation.

Figure 1



←More Progressive More Conservative→

Healthcare	🟦 Medicare for All	🟪 Restore ACA	🟥 Repeal ACA	🟥 Repeal ACA/ Privatize Medicare
Immigration	🟪 Amnesty/Path to citizenship	🟪 Permanent DACA	🟥 Repeal Asylum Law	🟥 Repeal "Birthright" clause
Tax Policy	🟦 2 % wealth tax	🟪 Higher marginal tax	🟪 Close Loopholes	🟥 Replace with national sales tax
Environment	🟦 Green New Deal	🟪 Rejoin PA + Carbon Tax	🟪 Cap + Trade	🟥 Restrict CA and other states
Safety Net	🟦 Guaranteed income	🟪 Expand earned income credit	🟥 Expand work requirement	🟥 Scale back Medicaid and Welfare
Voter Access	🟦 Voting by mail/early voting	🟪 Online/automatic voter registration	🟥 Require photo ID	

🟦 Democrat candidates only

🟪 Both parties

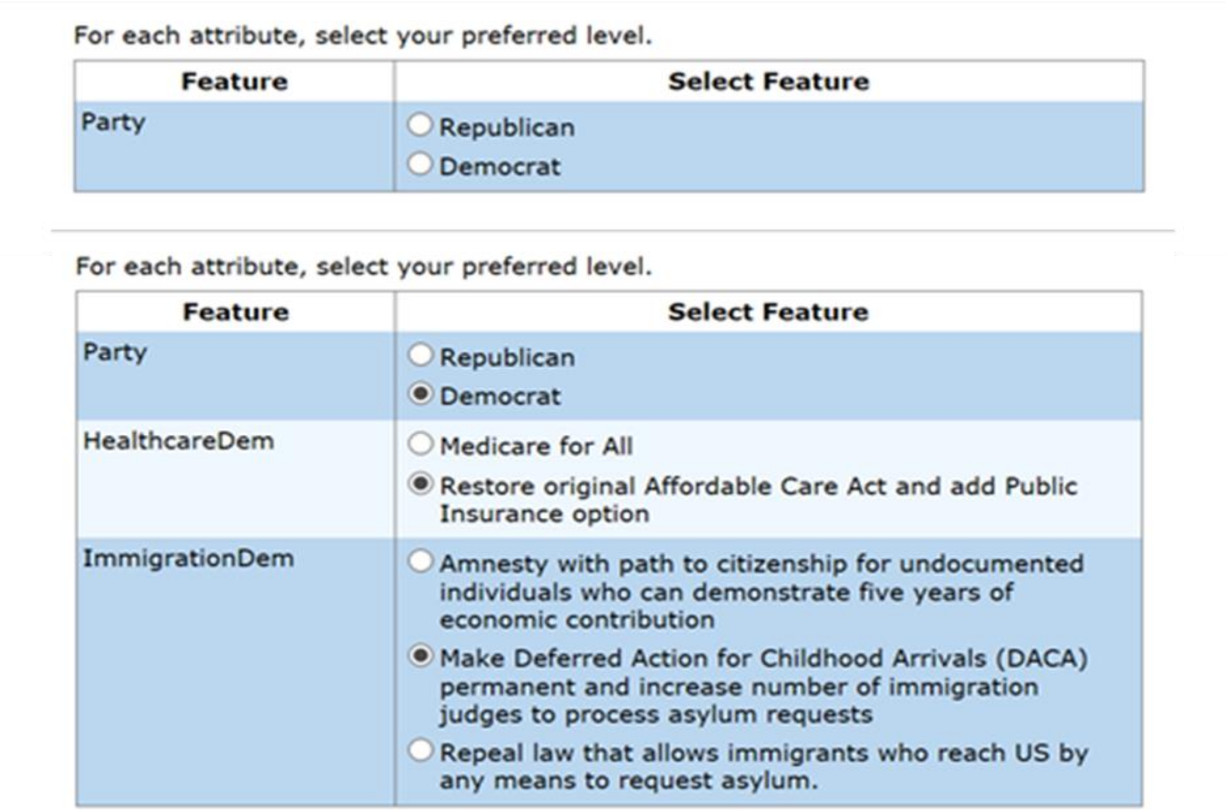
🟥 Republican candidates only

We placed no restrictions on the democratic values statements; candidates of either party could endorse both the more democratic and less democratic positions.

With an alternative-specific design in ACBC, the BYO exercise first asks respondents to pick a level for the primary attribute (in this case, candidate's party). The BYO exercise then displays only attributes and levels that are linked to that level of the primary attribute. Figure 2 shows how this would have looked in our study. Once the first question is answered, the entire BYO question appears with the conditional attributes (in this case, the attribute labels indicate that they apply to Democrat candidates). This presented a dilemma. We hoped to include all policy positions in the BYO exercise so that, for example, a respondent who might prefer a Republican candidate could still express a preference for "Medicare for all" or the "Green New Deal." We decided to move the policy preference questions out of the ACBC and skip the BYO section. We gave up the ability to use some prior information in the design of the screening concepts in order to get full coverage of the policy positions across all respondents. Figure 3 illustrates the way we asked these questions outside of the ACBC.[4]

[4] We recognize that there are other ways we could have approached this problem. For example, we might have asked for ratings on each of the policy positions and then used the constructed list capabilities in Lighthouse Studio to create BYO lists that reflected the policy positions that were acceptable to each respondent.

Figure 2



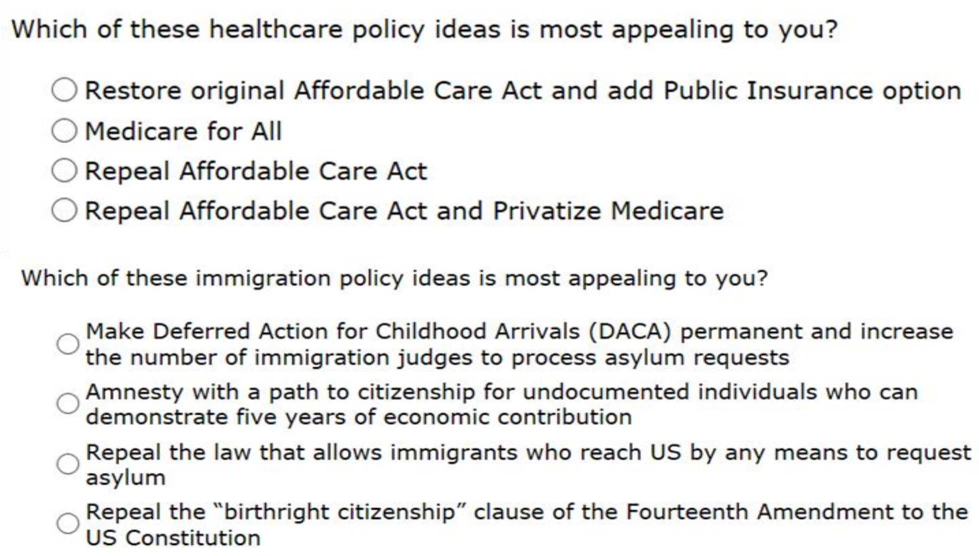
For each attribute, select your preferred level.

Feature	Select Feature
Party	○ Republican ○ Democrat

For each attribute, select your preferred level.

Feature	Select Feature
Party	○ Republican ● Democrat
HealthcareDem	○ Medicare for All ● Restore original Affordable Care Act and add Public Insurance option
ImmigrationDem	○ Amnesty with path to citizenship for undocumented individuals who can demonstrate five years of economic contribution ● Make Deferred Action for Childhood Arrivals (DACA) permanent and increase number of immigration judges to process asylum requests ○ Repeal law that allows immigrants who reach US by any means to request asylum.

Figure 3



Which of these healthcare policy ideas is most appealing to you?

○ Restore original Affordable Care Act and add Public Insurance option
○ Medicare for All
○ Repeal Affordable Care Act
○ Repeal Affordable Care Act and Privatize Medicare

Which of these immigration policy ideas is most appealing to you?

○ Make Deferred Action for Childhood Arrivals (DACA) permanent and increase the number of immigration judges to process asylum requests
○ Amnesty with a path to citizenship for undocumented individuals who can demonstrate five years of economic contribution
○ Repeal the law that allows immigrants who reach US by any means to request asylum
○ Repeal the "birthright citizenship" clause of the Fourteenth Amendment to the US Constitution

Unlike the Bright Line Watch study, we did not include any candidate characteristics in the conjoint experiment. Many salient candidate characteristics such as perceived competence, integrity, and compassion are difficult to operationalize. Hainmueller et al. used attributes that might be considered correlates of competence and morality, such as profession, religion, and level of education, but the levels they selected (especially considering the context of a presidential election) may not be the same cues that voters use to assess competence, integrity, or compassion. Finally, to the extent that we wish to predict possible election outcomes, we may not be able to represent the actual clusters of traits presented by any real set of candidates.

Nonetheless, we are in the run-up to the next U.S. presidential election and we wished to incorporate some information about the actual candidates who have announced their intention to pursue a party nomination. In addition to the ACBC exercise, we included an anchored MaxDiff with actual contenders for the 2020 U.S. Presidential election as the items. The anchoring question used a five-point voting likelihood scale. The objective of the MaxDiff was to obtain information on candidate preferences that we might use to model different voting scenarios, such as a Democratic party primary contest. We included the top ten Democratic contenders based on polling at the time the study was fielded, as well as President Trump and the one announced Republican challenger.

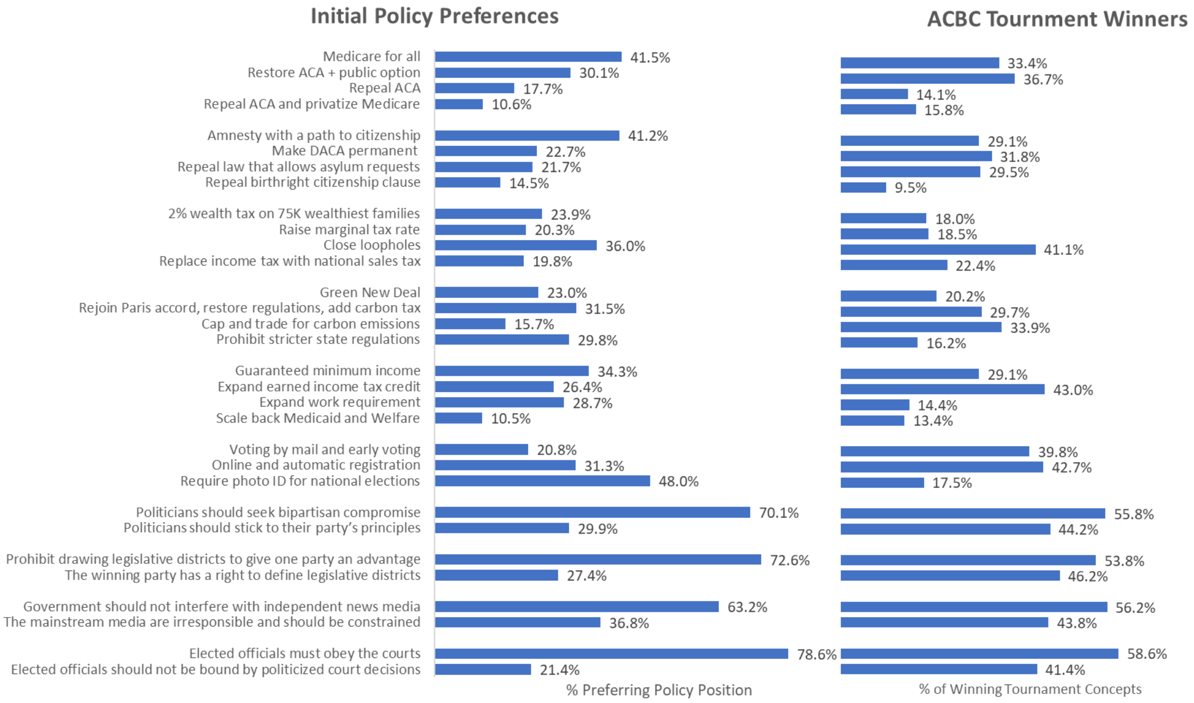
ACBC Results

In addition to the estimation of utilities, ACBC provides a wealth of both aggregate and individual-level information about the choices that respondents make. These include the proportion of respondents who identify specific attribute levels as "must have" or "unacceptable," as well as tallies of the number of times each attribute level appeared in the winning profile from the tournament section of the exercise.

We found that about 14% of respondents identified one party as either a must have or unacceptable, with 7% saying that Republican was a must have/Democrat was unacceptable and a similar proportion saying that Democrat was a must have/Republican was unacceptable. Given that 28% of the sample self-identified as "strong Democrat" and 19% as "strong Republican," these results indicate that respondents may be less adamantly partisan than we might expect.

We compared the policy and value choices that respondents made in our pseudo BYO questions to the frequency with which those same policies and values were included in the tournament section winning concepts. These results are presented in Figure 4.

Figure 4



For the policy positions on key issues (healthcare, etc.) it appears that respondents are somewhat more likely to prefer more extreme positions when asked to select the one policy or value statement that is most appealing compared to their choices in the tournament section of ACBC. In part this is explained by the fact that in the tournament, some policy positions were only available if, say, the respondent chose a Republican candidate. For example, 41.5% of respondents (including 32% of Republican/Republican-leaning) prefer "Medicare for all" to the other healthcare policies, but there were no Republican candidate profiles that supported this policy. Similarly, 32.6% of Democrat/Democrat-leaning respondents favor requiring that voters show a photo ID for national elections, but no Democratic candidate profiles supported that policy.

Recall that one of our objectives was to determine the extent to which voters might trade off democratic values to make a partisan choice. Our results show strong endorsement of the four democratic values in the pseudo BYO responses. For example, 78.6% say that elected officials must obey the courts even when they think the decisions are wrong. In the tournament section, however, only 58.6% of winning concepts included this value statement. Unlike the policy positions, the democratic values statements were not conditioned on the candidate's party, so a Republican profile would be just as likely to endorse obeying the courts as would a Democrat profile.

A sensitivity analysis based on the individual-level utility estimates shows that party affiliation has the biggest impact on candidate choice. For the policy attributes, most policy effects are small but of an order of magnitude that might be enough to make a difference in an election. The impact of policy statements depends on a candidate's party affiliation. For Democratic candidates, endorsing Medicare for all leads to a 3% increase in predicted preference share compared to restoring the Affordable Care Act and adding a public option. For Republicans, endorsing repeal of the ACA leads to a 3% increase in predicted preference share compared to restoring the ACA and adding a public option. Immigration policy is interesting. This is the only policy attribute where a Republican candidate could endorse the extreme left position ("Amnesty with a path to citizenship," which was the core of previous bipartisan legislation that failed to pass in Congress). Both Republican and Democrat candidates are slightly rewarded for endorsing amnesty with a path to citizenship, and candidates of both parties are penalized to varying degrees for more restrictive immigration policies. In addition, candidates of both parties are penalized for endorsing undemocratic value positions. Figures 5 and 6 display the results of the sensitivity analysis.

Figure 5

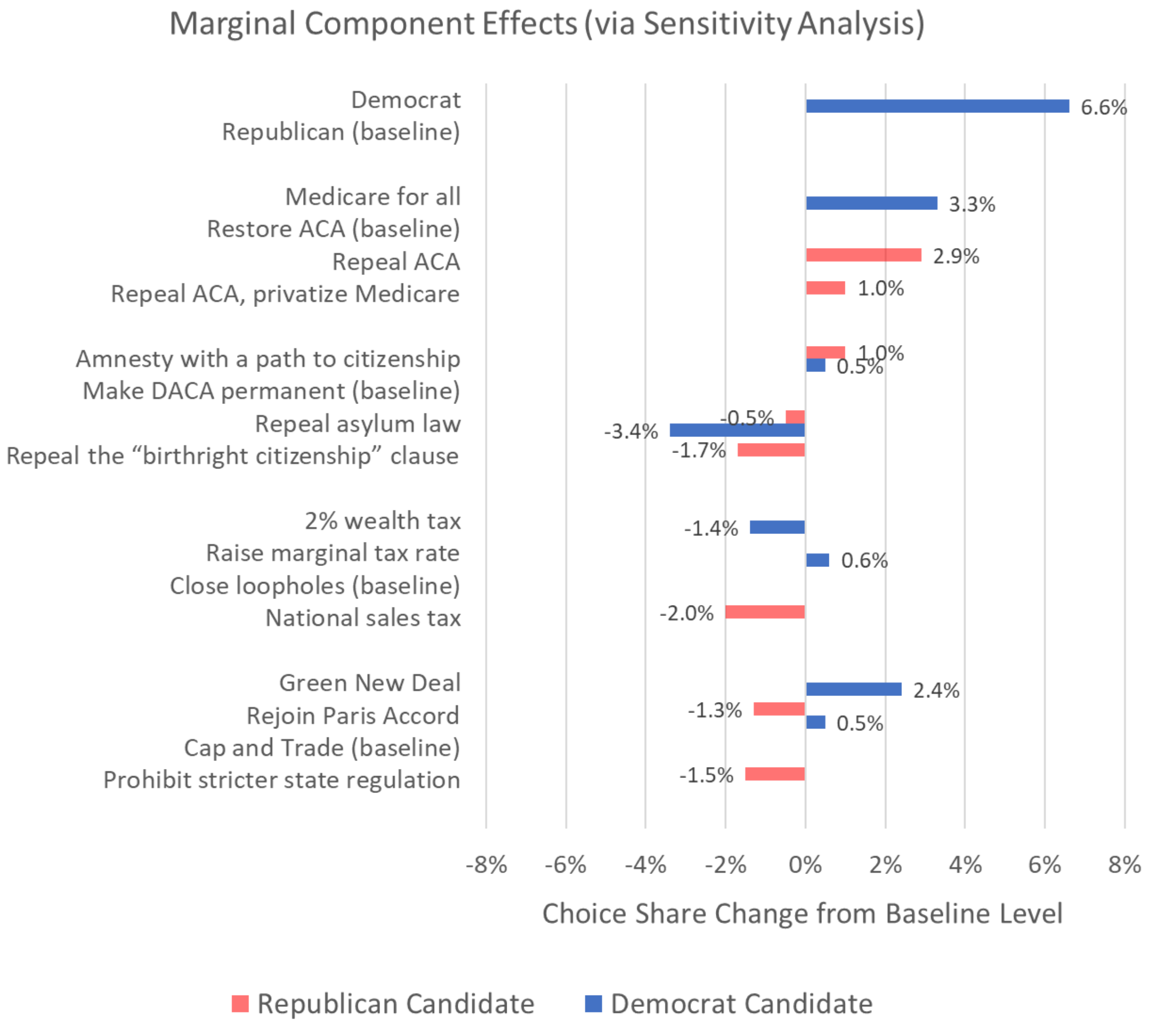
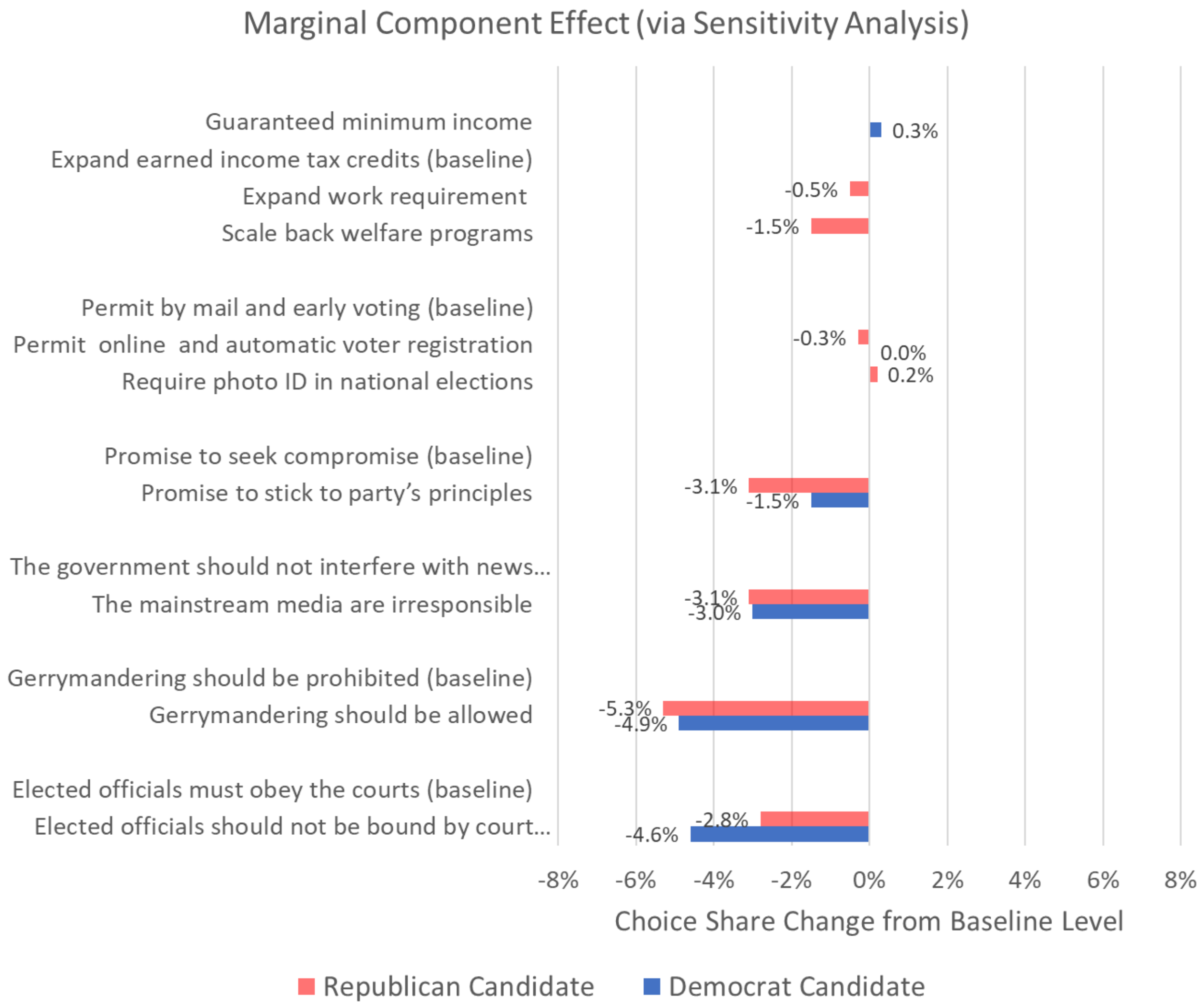
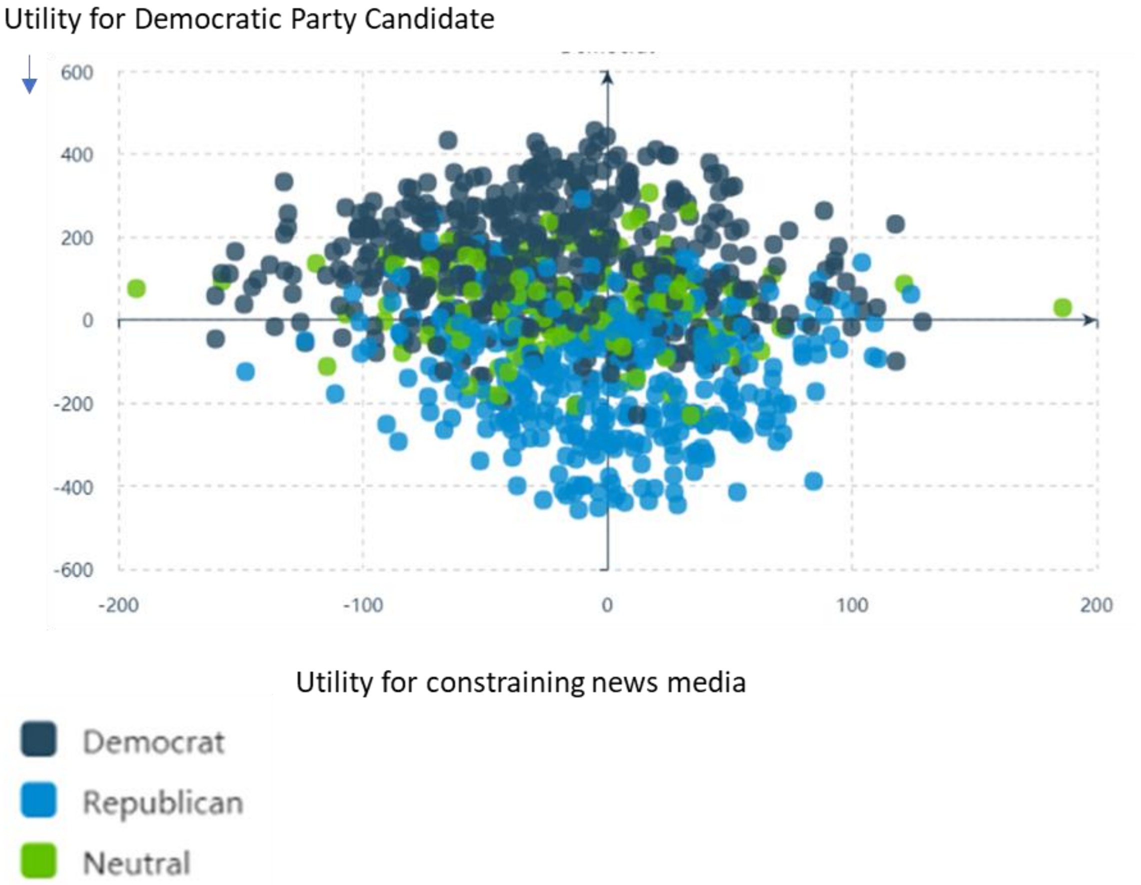


Figure 6



One of the potential advantages of a disaggregate model of voter preferences, whether we obtain it by ACBC, CBC with hierarchical Bayes estimation, or some other disaggregate method, is the ability to explore the relationship between the estimated utilities and other respondent characteristics. In particular, we are interested in the distribution of policy and party preferences for three distinct groups of voters: Democrats (strong Democrats, not very strong Democrats, and Democrat-leaning independents), Republicans (strong Republicans, not very strong Republicans, and Republican-leaning independents), and all other "neutral" voters (either do not prefer any party or prefer a third party). Scatterplots provide a simple look at the relationship between preferences and party affiliation. Figure 7 presents one such scatterplot showing the relationship between utility for a Democratic party candidate and the utility for the undemocratic value statement that the news media should be constrained. The plot indicates that Republican-leaning and neutral respondents are more likely to have a positive utility for constraining the news media. Those Republican-leaning respondents who have positive utility for a Democratic candidate are about equally likely to have negative utility for constraining the media as to have positive utility for this undemocratic value statement.

Figure 7: Scatterplot of Utilities (Zero-Centered Differences) for Democrat vs. Constraining New Media



Such plots might be useful to candidates or campaigns who want to know which policies and positions would appeal both to "base" voters as well as those who are neutral or lean towards a different party.

Candidate MaxDiff Results

The twelve named candidates were presented in sets of four; the MaxDiff question was:

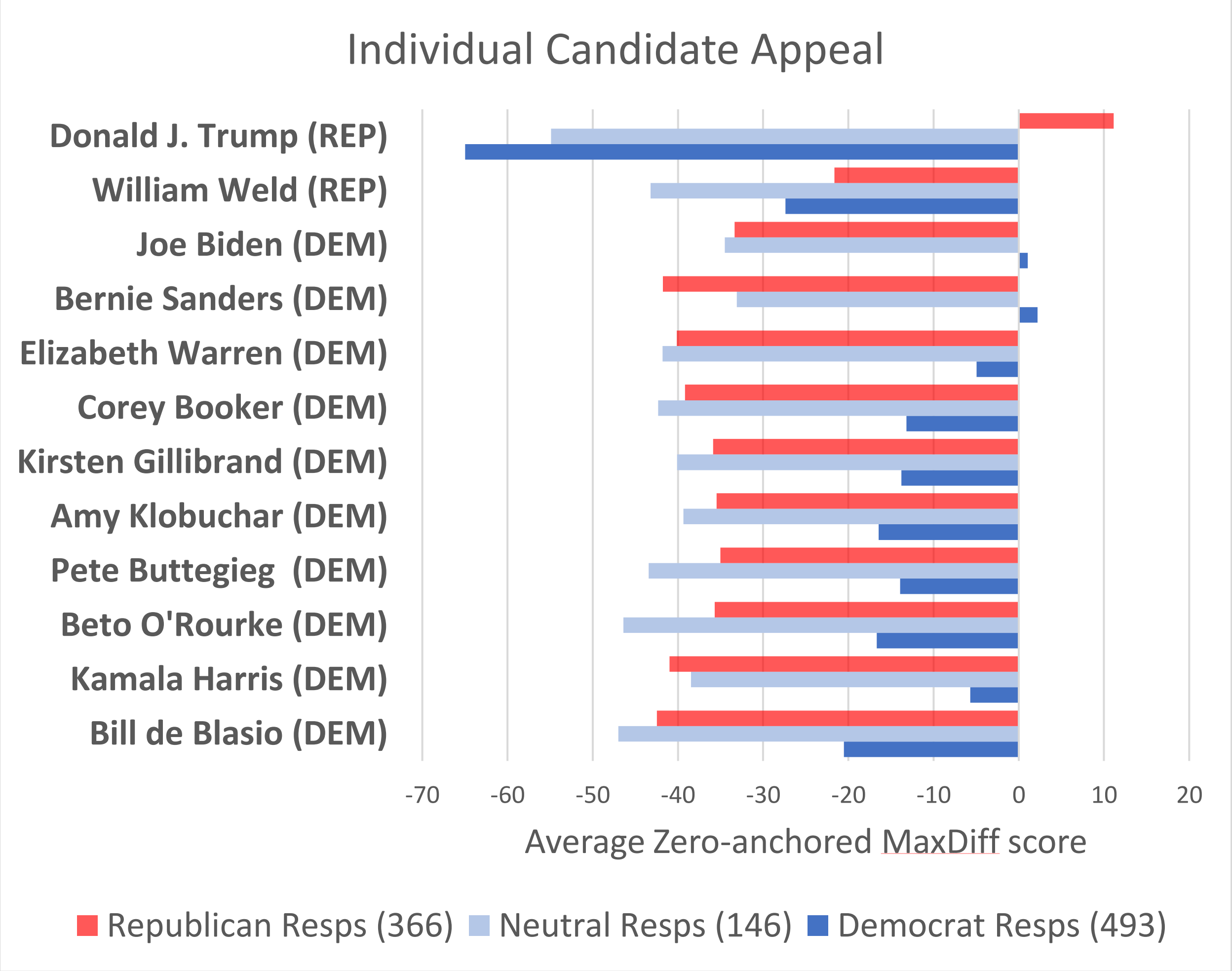
> "Considering only these four declared presidential candidates, which one are you most favorable towards, and which one are you least favorable towards?"

The anchoring question was:

> "How likely are you to vote for each of these candidates if they are their party's nominated presidential candidate?"

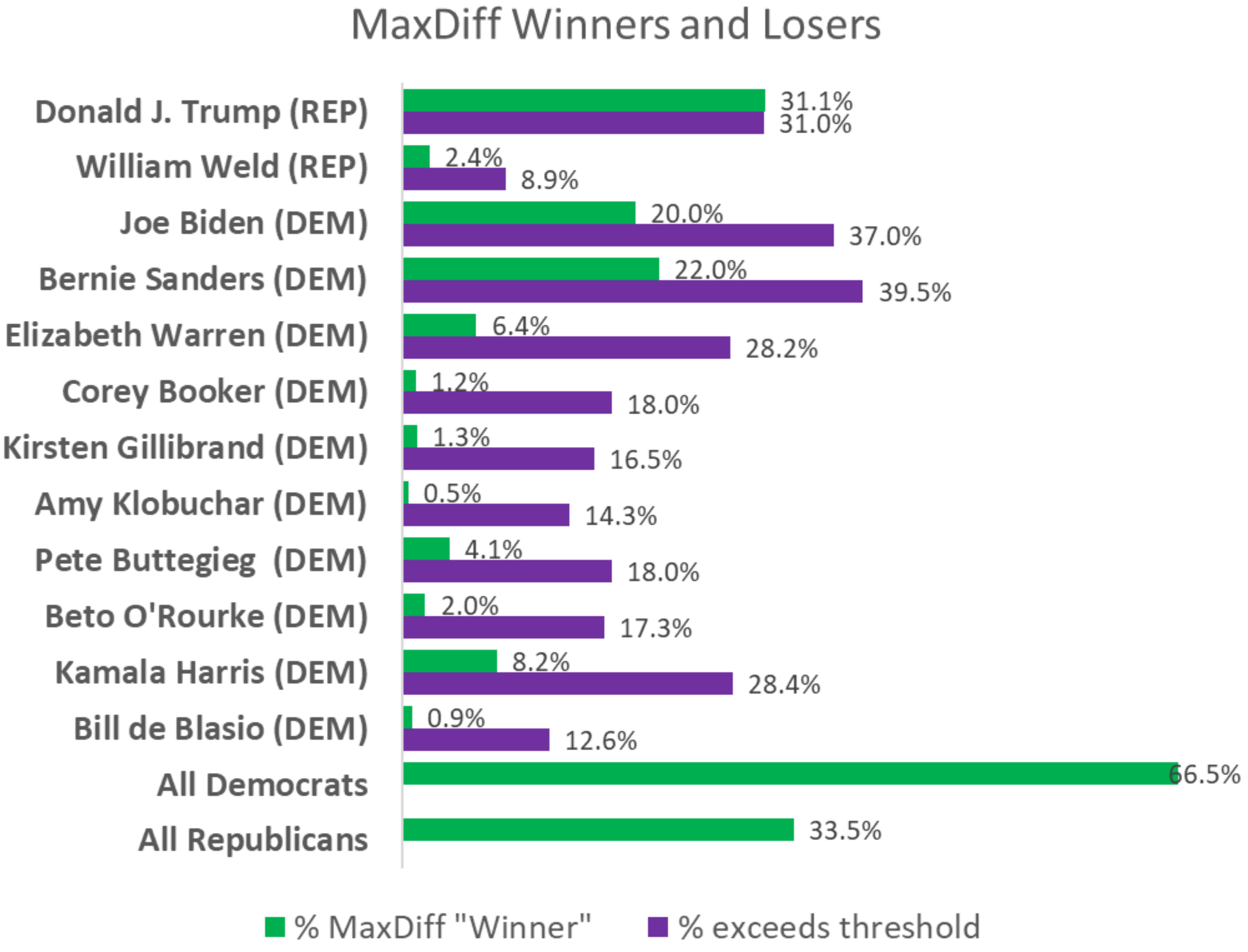
We used a five-point likelihood scale ranging from "Definitely would not vote for" to "Definitely would vote for." We dichotomized the scale, with "Probably" and "Definitely will vote for" responses treated as exceeding the threshold. Figure 8 shows the zero-anchored MaxDiff results. On average, only three candidates exceed the anchoring threshold: President Trump, former Vice President Joe Biden, and Senator Bernie Sanders. This probably reflects the public's uncertainty about the many candidates seeking the Democratic party nomination.

Figure 8



We also looked at the number of times each candidate was the "winner" of the MaxDiff exercise, and the number of times each candidate exceeded the zero-anchor threshold. Figure 9 displays these results. In total, Democratic party candidates win the MaxDiff 66.5% of the time, while President Trump wins 31.1% of the time. President Trump also exceeds the zero-anchor threshold 31% of the time, indicating that his support is consolidated. Among Democrats, Senator Sanders has the highest win rate (20%) and exceeds the threshold most often (39.5%) but former Vice President Biden is a close second. Overall, these results suggest that voter preferences had not coalesced around a single "best" candidate at the time the study was fielded.

Figure 9



Integrating the ACBC and MaxDiff Results

We wished to use the ACBC model to simulate electoral outcomes based on actual candidates. While we can create hypothetical candidates that match, at least to some extent, each candidate's policy positions, we would lack information about the excluded candidate characteristics. On the other hand, the MaxDiff contains information about the overall value or appeal of each candidate's constellation of attributes but does not directly capture the appeal of specific policies.

We devised two different "naïve" approaches to integrating the MaxDiff utilities into a market simulator with the ACBC utilities. In the absence of a common anchoring attribute, a major obstacle to integrating MaxDiff utilities with ACBC utilities is scale factor difference. Utilities have no absolute value, and each respondent can have a unique scaling of the utilities. With a common set of parameters, such as the ACBC utilities, the individual utility estimates can be transformed to zero-centered differences to eliminate the scale differences between individuals. However, there is no guarantee that such a transformation will put the MaxDiff and ACBC utilities on a common scale.

If we define our objective in integrating the two sets of utilities as creating a situation where, for two named candidates from the same party with identical policies, the probability of choosing one candidate over the other is a function of the differential appeal of the candidates, we could look at the ratio of the MaxDiff utilities for the two candidates without any rescaling. However, if we have two candidates with different policies, scale differences between the two sets of utilities will distort the predicted choice probabilities.

For our first naïve approach, we used the raw utilities from the ACBC and MaxDiff models. We rescaled the MaxDiff utilities so that the total range of the MaxDiff utilities was equal to the largest utility range for any single attribute in the ACBC model, which in all cases was the candidate's party affiliation. We then added the rescaled MaxDiff utility for each candidate to the sum of ACBC policy-based utilities (prior to exponentiation or any other transformation).

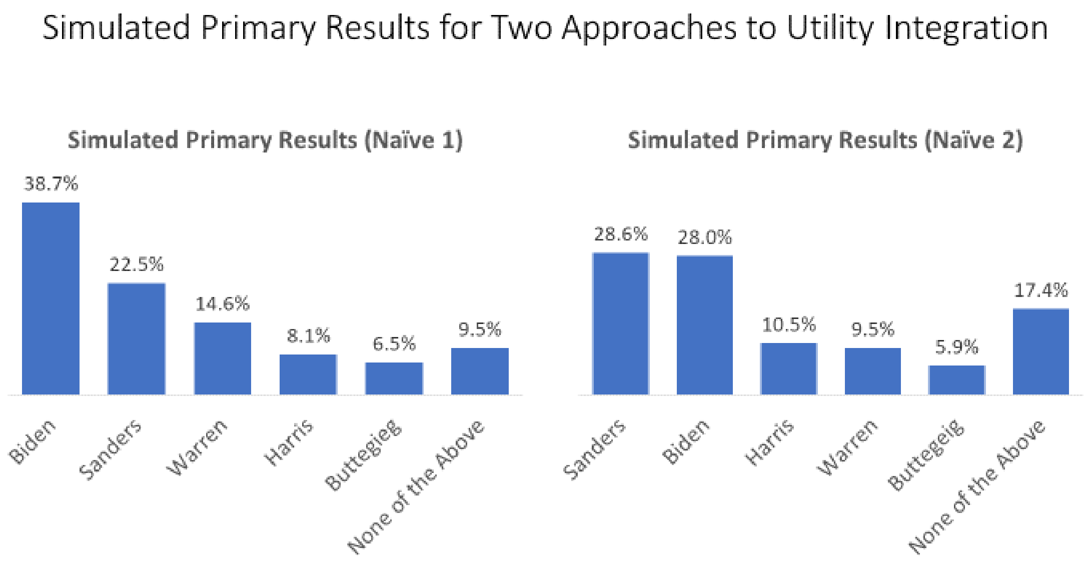
For the second naïve approach, we added the zero-centered differences transformation of utilities to the zero-anchored scores from the MaxDiff and calculated the share of votes using a simple first choice rule.

Scenario Simulations

We ran market simulations using both integration approaches. We first simulated a Democratic primary contest with the five top-polling Democratic candidates (Biden, Sanders, Warren, Harris, Buttegieg). We matched the policy positions as closely as possible to those endorsed by the candidates. For example, both Sanders and Warren proposed Medicare for all, while the more moderate candidates endorsed restoring the Affordable Care Act and adding a public option.

Figure 10 compares the results of simulations with the two different integration approaches. Biden does much better with naïve approach one than under naïve approach two. Naïve One is influenced both by ordinal preference and the difference between candidates while Naïve Two is influenced primarily by the ordinal preference. Biden and Sanders were winners or exceeded the threshold roughly the same number of times in the zero-anchored MaxDiff results.

Figure 10



We also simulated head-to-head contests between President Trump and former Vice President Biden and between President Trump and Senator Warren. Figure 11 shows the simulation results for the first naïve integration approach, and Figure 12 compares the simulations for the second naïve integration approach. With a first-choice rule for the simulations, the results between the two methods are very similar.

Figure 11

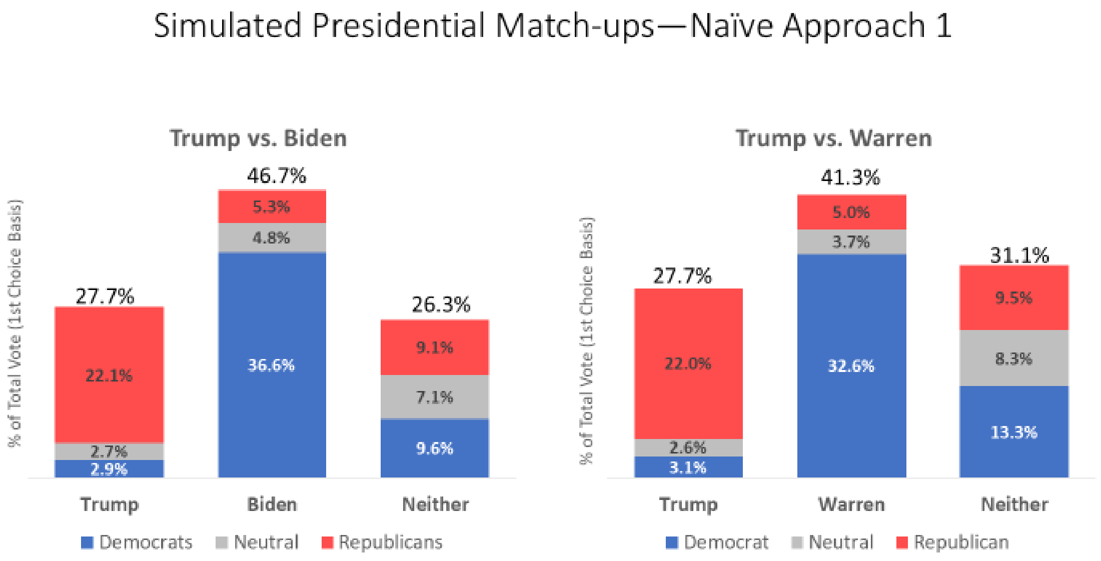
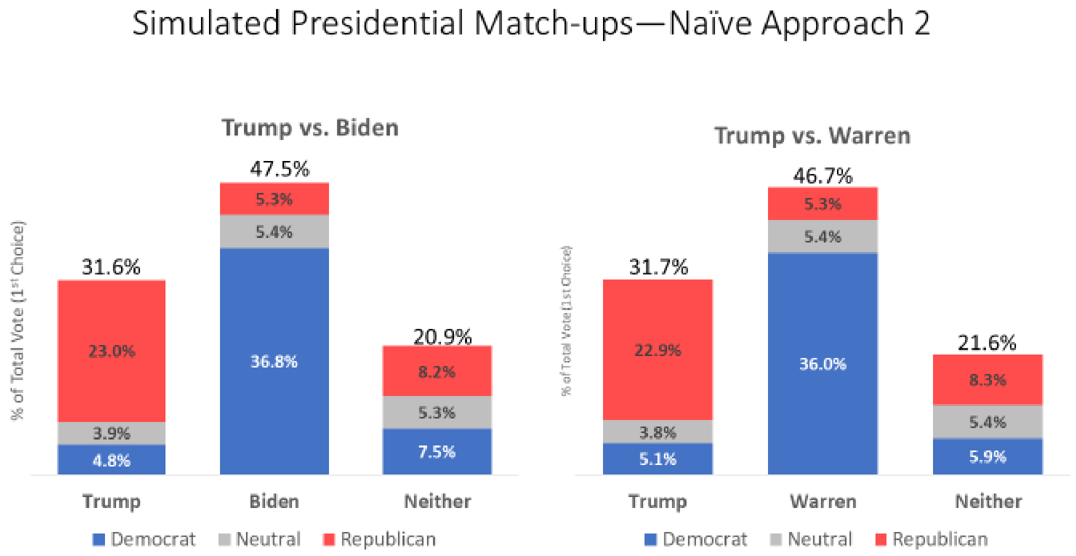


Figure 12



Conclusions and Implications

We sought to answer two questions with this study:

- Is ACBC an appropriate and perhaps better approach than other conjoint methods for understanding voter preferences and predicting their electoral choices?
- To what extent will voters trade off democratic values in order to maintain partisan loyalty.

On the first question, we conclude that *disaggregate* estimation of conjoint utilities (in our case, hierarchical Bayes) provides more insight into the distribution of voter preferences than the aggregate estimation methods employed by previous applications of conjoint analysis in political science. In our view, HB estimation is a bridge between sociological and psychological approaches to understanding voter behavior. Sociological theories seek to explain voting in terms of socio-demographic factors and group identity, while psychological theories focus on *intrapsychic* factors. HB turns out to be a powerful tool for linking individual preferences to external variables. Yang and Allenby (2003), for example, developed a model of interdependent consumer preferences and found that preferences for Japanese automobiles are related to geographically and demographically defined networks.

As implemented by Sawtooth Software, ACBC captures data that provides a detailed look at the choices made by respondents. Of particular interest to our study was information about the frequency with which a candidate’s party affiliation was a “must have” (or “unacceptable”) factor and the correspondence between “build your own” policy choices and the frequency with which those choices were included in each respondent’s winning concept in the tournament section of the ACBC.

That being said, the unique advantages of ACBC over standard Choice-Based Conjoint (with HB estimation) for understanding political preferences are small, at best.

With respect to predicting voting behavior, our ACBC model incorporates only some of the variables that could influence candidate choice. We included a few policy positions and party affiliation. Actual choices will depend on candidate characteristics as well. We measured overall preferences for specific real candidates with a MaxDiff exercise without any attempt to assess which candidate characteristics drive those preferences. We attempted to integrate these candidate preferences with the policy and partisan preferences from the ACBC. However, in the absence of any bridging attributes between the ACBC and MaxDiff exercises, any decisions we've made about the magnitude of the scale differences is arbitrary. At best, we can say that the MaxDiff utilities can serve as a tie breaker for any pair of candidates that have the same policy positions.

We think that for the next iteration of this research program, including candidates in the ACBC is desirable. Given the large field of Democrat contenders for the 2020 election, we might use some variety of MaxDiff (or some other screening criterion) to narrow the list of candidates for the ACBC or CBC exercise.[5] We could also capture perceptions of these candidates on the factors we identified in the introductory section of this paper to determine which characteristics drive preferences for individual candidates.

On the second question, our findings are similar to those reported by Bright Line Watch. Our respondents do appear willing to trade off their preferences for democratic values in order to choose a candidate that reflects their policy and party preferences. We see this most clearly in comparing the preferences for democratic values expressed before they entered the ACBC and the proportion of times those preferred values were included in their tournament winning candidate profiles.

We believe that disaggregate Choice-Based Conjoint (CBC) analysis offers great potential for extending our understanding of voter preferences and the way they make electoral choices. Moreover, CBC seems well suited for refining policy positions on key issues. In particular, understanding the appeal of different policies among independents and those who prefer third parties could yield insights into the best ways to attract those voters.

David Bakken

Gretchen Helmke

Mitch Sanders

[5] Using MaxDiff to prune the list of candidates was suggested by Megan Peitz.

REFERENCES

Bartels, L. M., "Partisanship in the Trump era." Working paper, Center for the Study of Academic Institutions (2018), Vanderbilt University.

Campbell, A, P. E, Converse, W. E. Miller, and D. E. Stokes, *The American Voter* (1960). New York: John Wiley and Sons.

Carey, J. M., K. P. Clayton, G. Helmke, B. Nyhan, M. Sanders, & S. C. Stokes. "Who will defend democracy? Evaluating trade-offs in candidate support among partisan donors and voters." Bright Line Watch (2019). https://preprints-api.apsanet.org/apsa/assets/orp/resource/5d02becd39ef0400184a49fd/original/blw-party-donors.pdf

Graham, Matthew, and Milan W. Svolik. 2019. "Democracy in America? Partisanship, Polarization, and the Robustness of Support for Democracy in the United States." Working paper, available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract id=3354559.

Hainmueller, J., Hangartner, D., and Yamamoto, T. (2015) Validating vignette and conjoint survey experiments against real-world behavior. Proceedings of the National Academy of Sciences, 112 (8). pp. 2395–2400. ISSN 0027-8424DOI:10.1073/pnas.1416587112

Hainmueller, J., D. J. Hopkins, & T. Yamamoto. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* (2014), 22: 1–30.

Svolik, Milan W. 2018. "When Polarization Trumps Civic Virtue: Partisan Conflict and the Subversion of Democracy by Incumbents." Working paper, available at: https://pdfs.semanticscholar.org/4d2c/50628b3333c52e6f0c7488cae125a996b3f3.pdf.

Yang, Sha and Greg M. Allenby, "Modeling Interdependent Consumer Preferences," *Journal of Marketing Research* (2003), 40(3), 282–294.

THE CHALLENGE OF IDENTIFYING CAUSALITY IN OBSERVATIONAL DATA

RAY POYNTER
THE FUTURE PLACE/NOTTINGHAM TRENT UNIVERSITY

SUMMARY

There has been an explosion in the amount of observational data available to decision makers and research. This growth, from social media, to transactional data, to passive tracking, presents exciting new opportunities to evaluate human behavior as it happens, rather than in the context of artificial experiments. At the same time, fields such as neuroscience and Behavioral Economics have been generating an increasing amount of concern about the veracity of data collected via questions (for example via surveys and focus groups). These two forces are driving the growth in the use of observational data to produce analytic models, predictive models, with the eventual aim of producing prescriptive models. However, there are challenges in the use of observational data, such as stepping from correlation to causality, survivor bias, homophily, and combinatorial effects.

This paper highlights the problems that can occur with observational data and offers potential solutions.

INTRODUCTION

We live in interesting times. There is an exponential explosion in the amount of data available. In an increasingly connected world, people are leaving a digital trail behind them. This observational data allows researchers to examine real life to understand human behavior. The shift from surveys and focus groups to observational data is a shift from what Christian Madsbjerg (2017) describes as the zoo to the savannah.

At the same time as the explosion in data is going on, there are growing concerns about the ability of people to explain their own motivations, beliefs, and intentions. Work by neuroscientists such as Antonio Damasio (1994) and Behavioral Economists such as Daniel Kahneman (2012) have illustrated that people struggle to report accurately on their own motivations and intentions. In addition, researchers have known for years that people's ability to recall events accurately is flawed.

These two forces, the growth of observational data and the concerns about asking questions to people, are the driving force behind the growing interest in using observational information to replace questions. The analysis of observational data is becoming the key method to understanding what people do, why they do it, and what they might do in the future.

The growth in the availability of observational data, for example social media, mobile phone data, web tracking, and loyalty card records, provide a wide range of exciting opportunities to explore human behavior. The use of this data is being employed to produce:

1. Descriptive analytics—which focuses on associations.
2. Predictive analytics—which makes the step to causality.
3. Prescriptive analytics—which seeks to attribute causes and to suggest optimal strategies to achieve specified aims.

While recognizing that these data sources present many opportunities to examine behavior and motivations, there are several key challenges that need to be acknowledged and, where possible, ameliorated. Among all the challenges presented by observational data, perhaps the key one is assessing causality. For data to be really useful, researchers need to move beyond descriptive analytics (which tends to focus on associations) to predictive and prescriptive analytics, which require assertions about causality.

This paper outlines some of the challenges in using observational data and suggests remedies and ameliorative measures.

OBSERVATIONAL DATA

There are many different sources of observational data and these sources produce a variety of types of information. The table below illustrates the breadth of different sources and implications.

Description	**Examples**	**Implications**
Big data	Bank records & Social Media	The need to address IBM's *Four V's of Big Data*, Volume, Velocity, Variety, and Veracity (IBM Infographic)
Census or samples	All phone records versus the records for Pay as You Go customers	Different rules for producing and using inferences
Objective or subjective	Till receipts versus ethnographic observations	Quantitative versus qualitative assessments
Structured or unstructured	Bank transactions versus uploads to Instagram	Paucity of tools for analyzing unstructured information
Behavioral or motivational	Loyalty card data versus emotions assessed by automated facial coding	Different measurement paradigms being used for behavior and motivations—often with different certainties
Naturally occurring or from experiments	Browsing data versus browsing data when A/B testing is employed	Need to assess whether other factors could be impacting the experiment
Observational only or observations with questions	Tracking people's movements via their phone versus tracking their movements and then asking them to describe the journeys	With observations the researcher has to make assumptions about the motivations, questions can provide additional information

The Opportunities Created by Observational Data

The interest in observational data is driven by two key forces. The first trend is the growing abundance in observational data, including big data, social media, and passive tracking. The second trend is a growing awareness of the limitations of questions as a method of understanding the world. Key examples of the limitations of direct questions include people's flawed memories and people's inability to access their own motivations. For example, researchers into eyewitness evidence in court cases have highlighted issues such as "believing is seeing" and "memory is malleable" that can render personal recall unreliable (Albright, 2017).

The Need for Causality

Historically, statistics has tended to avoid making assertions about causation, preferring to focus on association and correlation. Indeed, one of the founders of statistics and the creator of the most widely used correlation coefficient, Karl Pearson, dismissed the pursuit of causation as "another fetish amidst the inscrutable arcana of even modern science" (Pearson, 1911). The trope "Correlation does not imply Causality" is one of the few statistical phrases that is in common usage, a phrase which seems to dismiss the need to properly investigate the underlying causes of the correlation.

The arrival of big data and algorithmic solutions led to some people questioning the need for scientific theories and assessments of causality. For example, Chris Anderson (Editor-in-Chief of Wired Magazine) said "There is now a better way. Petabytes allow us to say: 'Correlation is enough'" (Anderson, 2008). This pronouncement created a storm of protest, perhaps best summed up by Nate Silver who described this view as ". . . badly mistaken." He went on to state "The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning" (Silver, 2012).

As mentioned above, a key driver for the use of observational data is to create predictive and even prescriptive models. Explanation, prediction, and prescription require causation (Watts et al., 2018) and so the balance of interest has shifted from mere association to causation. Indeed, some describe it as more than a shift, observing that the opportunities, needs, and new approaches have "spawned a revolution in the way causality is treated*"* (Pearl et al., 2016).

Potential Problems with Observational Data

There are a wide range of challenges presented by observational data and this paper will discuss a few of them, in order to highlight the sorts of issues that need to be addressed. However, this paper should in no way be taken as a recommendation against using observational data. One description of observational data is "No Questions" research. The 2019 ESOMAR Global Market Research Report suggested that nearly 50% of all market research (in global dollars spent) was conducted via No Questions research (Poynter, 2019). Furthermore, No Questions research is growing rapidly, unlike traditional Questions Research (e.g., questionnaires and focus groups) (ESOMAR 2019). Observational data should be embraced and utilized, but the challenges should be recognized and dealt with.

The potential problems covered in this paper are:

- Spurious correlation
- When observations offer the wrong message
- Combinatorial effects
- Ignoring the true driver
- Complex and/or chaotic relationships
- Observer and/or measurement effects
- Survivorship bias
- Not explaining the why
- Why are economists so bad at predicting recessions?
- Confusing influence and homophily
- Observational data and the rear-view mirror

Spurious Correlation

Correlations can be highly seductive, especially if the correlations appear significant and plausible. One key problem with correlations are what are referred to as spurious correlations, correlations that are the result of pure chance. As data becomes large "the overwhelming majority of correlations are spurious" (Calude & Longo, 2017). There is a website devoted to highlighting spurious correlations (https://www.tylervigen.com/spurious-correlations), highlighting correlations such as the correlation of 0.9979 between annual US spending on science, space, and technology and the annual number of suicides by hanging, strangulation, and suffocation between 1999 and 2009.

When there is a genuine correlation, researchers should seek to find out what is causing the correlation. When there is a genuine and persistent correlation between X and Y the connection could be: X causes Y, Y causes X, both X and Y are caused by some other factor Z, or X and Y are connected via a feedback loop where they both influence each other.

When the correlation is spurious, researchers need to identify it as spurious and counsel against its use in decision making. Amongst the things that help identify a correlation as spurious are:

1. Is there a line of communication between the two events? In the example earlier between US spending on space and suicides, there is no apparent line of plausible influence from one to the other.
2. Was the correlation found because a link was being investigated, or because a large data set was trawled looking for associations? If an association is found through trawling a large data set, the chance of it being spurious is much higher.
3. Later in this paper, I will discuss causal inference. Some of the techniques in causal inference (such as causal graphs) can be employed to help detect spurious correlations.

When Observations Offer the Wrong Message

A good example of an observational study, initially, offering the wrong message is provided by studies in 2012 that suggested running was bad for people's hearts. The key study was research by Duck-Chul Lee with 50,000 patients (presented at the American

College of Sports Medicine 59th Annual Meeting, 2012), along with heart findings from James O'Keefe that looked at issues such as fibrosis, calcified arteries, and arrhythmias (O'Riordan, 2012). The news that running appeared to damage hearts was picked up by the media, who published numerous stories about how running was bad for you, and that anything other than a small amount of exercise was either useless or damaging.

Subsequent analysis showed that the conclusions, relating to the impact of running on people's hearts, were based on flawed analysis. The key issue with Lee's study was identified by Thomas Weber (2013). The sample of 50,000 people contained some long-distance runners, including marathon runners. Lee wanted to compare these people with the non-runners and the occasional runners. To remove sources of bias, factors such as race, age, and gender were controlled for. But the research also controlled for factors such as body mass index, blood pressure, and cholesterol levels. The reason for controlling for these variables is that these are risk factors for cardiac-related issues. However, regular running and running longer distances tends to change these numbers, it tends to reduce weight, lowers blood pressure, and brings cholesterol levels down. By controlling for these factors Lee had changed the results in the wrong direction.

When the data were re-processed by Lee and published in a peer-reviewed journal, without controlling for these correlated characteristics, the message changed (Lee, 2014). In the new analysis running longer distances was no longer a negative. The main thrust of the new paper was that running at least 5-10 minutes a day offered "dramatic reduction" in deaths from heart disease.

These sorts of problems can be ameliorated by applying techniques such as causal inference to help determine which factors should be controlled for and which should not, to help the researcher match people in one group with appropriate people in other groups. These techniques help map out the relationships in the data and suggest the best way of increasing the chances of finding causal links.

Combinatorial Effects

The impact and challenge of combinatorial effects is perhaps best shown with a disguised and simplified example from an advertising campaign from the UK. Start by assuming that in a region of the UK the following scenario is conducted.

Region A

- In time period 1, the sales were indexed to 100.
- In time period 2, a TV advertising campaign is run and the sales go to 110.
- In time period 3, a Twitter campaign is added to the TV campaign. The indexed sales go to 130.

The implication would appear to be that TV increases sales by 10 percentage points, and Twitter increases the sales by 20 percentage points.

However, in other regions, different patterns were observed.

Region B

- In period 1, the sales were indexed to 100.
- In period 2, a Twitter campaign is run and the sales go to 110.
- In period 3, a TV advertising campaign is added to the Twitter campaign. The indexed sales go to 130.

Now the implication appears to be that combining Twitter and TV create a lift of 20 percentage points, compared with running just one of the campaigns. But this insight is only available if the observational data is capable of being broken into these two cases.

The model can be extended with a third region, a region where no advertising was run.

Region C

- Period 1, sales were indexed to 100.
- Period 2, sales were 105.
- Period 3, sales were 110.

The picture that now emerges is that some of the changes seen in Region A and Region of B would have happened anyway. Region C is the counterfactual—what would happen if there were no advertising.

The best way of determining combinatorial effects is to design campaigns so that they can be measured properly. The IPA in their report The Expert Guide to Measuring Not Counting described this process as baking the measurement into the campaign (IPA, 2015). In the absence of other data, when a treatment A is added to a treatment B, it should be assumed that the final effect C is given by f(A) + f(B) + f(A&B) + U. Where f(A&B) is the combinatorial effect, and U is an umbrella term covering unknown factors and measurement error.

Coverage Error

If we look at the beach at low tide, we will not understand what it looks like at high tide. If we measure traffic flows during the weekend, we will not understand the flows during the working week. If we just analyse the shopping behavior of people with a loyalty card, we will not understand the total picture, including those who do not have a loyalty card. All of these are examples of challenges created by coverage error.

A good example of how coverage errors can change the results in an observational study was provided by a transport study in Germany (Gruschwitz & Schönduwe, 2017). The paper reports on a long-standing study in Germany that explored journeys. The study used a sample of people collecting travel diaries to measure their journeys. The concern with the diary approach was that people tend to forget some trips, and they may simplify their journeys (for example describing a 55-minute journey as lasting one hour). The researchers provided a sample of citizens with an app to load on their mobile phone. The app recorded journeys and the participants could add extra details later.

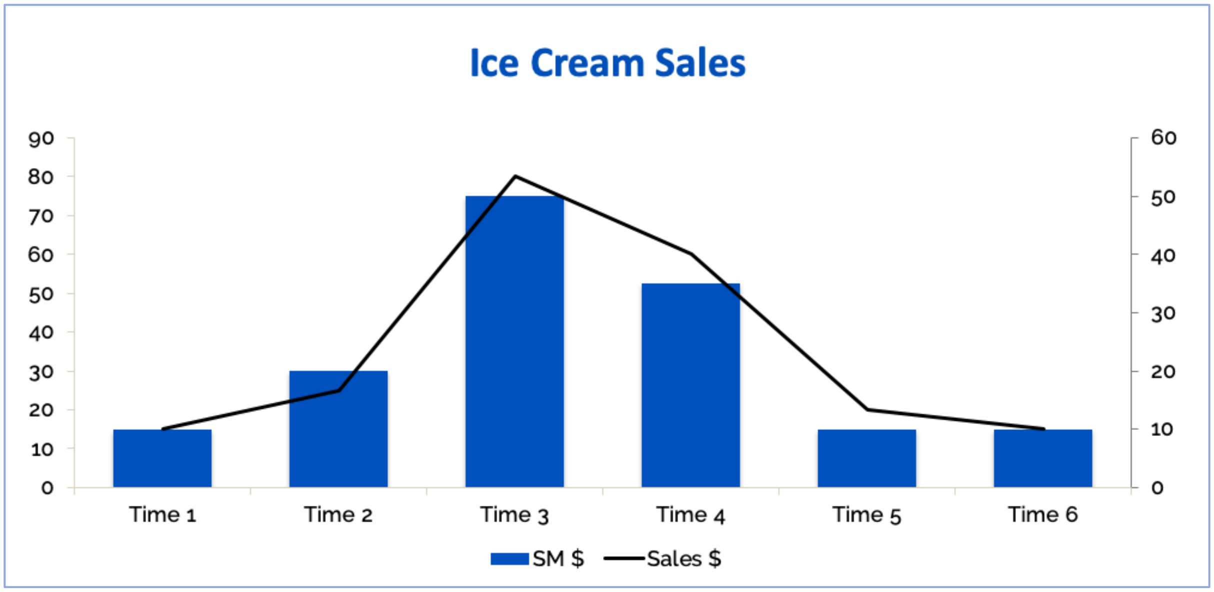
When the researchers analysed the data, they found some good news and some bad news. The app seemed to work and journeys did not show "heaping" (being rounded to 10KM, or 20 minutes). However, the data recorded 16% fewer journeys (with 11% less distance being travelled and 18% fewer minutes). This was a surprise, the researchers had expected more journeys, not fewer. When the researchers reviewed the research, they noted that there were holes in the data. The travel app was programmed to turn itself off when a phone's battery charge fell to 20%, an event that turned out not to be a rare event. The problem with the observational data was a coverage error, the data did not cover periods where phones were less charged, an event that happens to most people some of the time, and some people a lot of the time. Once the researchers were aware of the problem, they were able to start finding remedies.

Whenever a researcher uses observational data, one of the key questions they should ask is "What is missing?" For example, what people might be missed and what situations might be missed?

Ignoring the True Driver

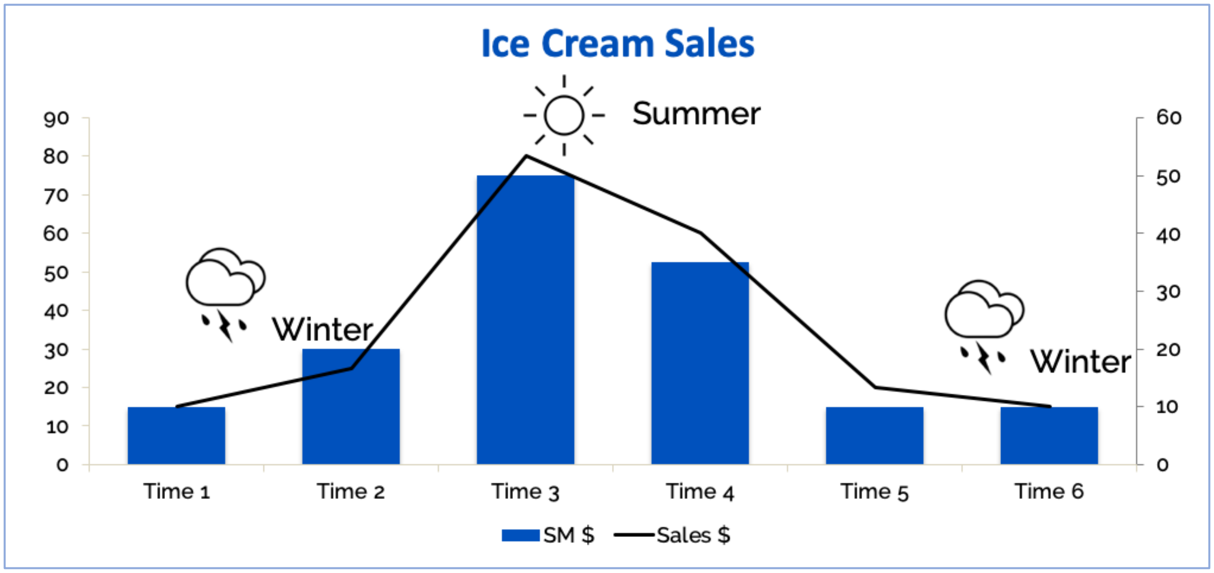
One of the uses that observational data is often put to is the determination of what events are "driving" some key outcome, in particular what events are driving purchase. The traditional market research route of asking people what caused them to buy X, or what might cause them to buy Y in the future, are fraught with problems, so the observational data seems a natural improvement. However, simplistic approaches to measuring drivers are also often flawed. Two key types of error are a) assuming that the last action before the event caused the event, and b) that the experiences that are measured include the actual driver of the event.

The two diagrams below illustrate this challenge. The diagram is an anonymized example from a leading CPG company and relates to a report from an insight professional. This first chart shows the trends in ice cream sales and the trends in social media advertising spend.



As the advertising spend increases the sales appear to increase. As the spend on social media campaigns fall, the sales fall. The result was that the person managing social media advertising called for more money, to keep spending at the higher level.

The insight professional took out his Sharpie and added the sketches shown on the diagram below.



What the second diagram shows is that the seasons determine the total sales of ice cream. Social media advertising might be shifting the market share, but not people's deep, seasonally linked behavior.

The starting point in avoiding the wrong driver trap is to keep in mind the fallacy highlighted by the Latin motto 'Post hoc, ergo propter hoc'—because B follows A, it is caused by A. The second key is to identify the counterfactual, what would have happened without the action being taken. In the example above, what would have happened without the social media advertising. Modern attributional modelling seeks to create a counterfactual (for example by matching people who experienced the stimuli with "lookie-likes" who did not experience the stimuli).

Complex and/or Chaotic Relationships

There can be an assumption that if we only had enough data, we could predict every outcome. For example, with enough sales data we should be able to forecast future sales. However, in many cases this is a mirage. In his book *The Signal and the Noise,* Nate Silver draws a distinction between three types of problems, typified by baseball, weather forecasting, and predicting earthquakes (Silver, 2012).

Silver has shown that baseball is a remarkably predictable game. The quality of a team is to a large extent a function of how good its individual players are. Metrics such as a pitcher's game score provide good insight into the quality of the team. For example, the pitcher's game score is given by the formula $gameScore{=}47.4{+}strikeouts{+}(outs{*}1.5){-}(walks{*}2){-}(hits{*}2){-}(runs{*}3){-}(homeruns{*}4)$ (FiveThirtyEight, 2019). This process is the essence of much of Silver's early fame, before he acquired much more fame for his election predictions.

Silver contrasts baseball with weather forecasting. Weather forecasting is complex and some of the elements are chaotic. However, over the years weather forecasting has gotten better and better, as models have become better, more data has become available, and computers are more powerful. Silver highlights two useful benchmarks for forecasting the weather. Benchmark 1 is to assume that the weather on the target day will be the same as today. This is a good method for tomorrow, and much less good for, say, 30 days from now. Benchmark 2 is to assume that the weather on the target day will be the average of the last ten year's weather for that date. Modern weather forecasting is now capable of beating these benchmarks for a period of up to eight days in advance. One of the key steps in weather forecasting is to simplify the data, a process known as discretizing (Christensen, 2015). The earth's atmosphere is divided into cuboids, for example 10KM by 10KM areas with a height that might vary from a few hundred meters to a few kilometers.

By contrast, the forecasting of earthquakes, in terms of locations, timing and size, has barely improved over the last fifty years. The US Geological Society say

> "Neither the USGS nor any other scientists have ever predicted a major earthquake. We do not know how, and we do not expect to know how any time in the foreseeable future. USGS scientists can only calculate the probability that a significant earthquake will occur in a specific area within a certain number of years. An earthquake prediction must define 3 elements: 1) the date and time, 2) the location, and 3) the magnitude."

Whenever we are working with a complex system, we need to start with an awareness that the link between the observed data and the observed outcomes may not be solvable. There may be necessary data that is not available, and the relationship may be too complex or chaotic for modelling to work. In cases like this it might make sense to change the objective of the modelling to something more modest. For example, the shapes of waves are chaotic and cannot be predicted, but the tides can be modelled and predicted.

As in the case of weather forecasting, one key step with complex systems is to reduce the complexity of the input variables, for example transforming them from continuous to discrete, grouping them, and simplifying them.

Observer and Measurement Effects

When people are aware that they are being watched their behavior can change. This effect is often termed the Hawthorne Effect. The Hawthorne Studies were conducted in North America from 1924 to 1932 are were initially intended to measure improvements in factory management and processes. However, analysis of the data suggested that the changes in things like output were heavily influenced by the fact that people knew they were being measured/observed. It should be noted that the original Hawthorne Studies have been somewhat discredited in terms of their methodology and rigor (Hassard, 2012). However, the use of the term Hawthorne Effect to describe modifications in behavior as a result of being aware of being observed is independent of the purpose and conduct of the original studies.

The nature of observer effects is difficult to predict with one meta-analysis of 19 studies concluding

> "Consequences of research participation for behaviors being investigated do exist, although little can be securely known about the conditions under which they operate, their mechanisms of effects, or their magnitudes." (McCambridge et al., 2014)

Measurement effects refer to changes that we create through the process of measuring something. For example, if we want to know the temperature of a glass of wine, we might put a thermometer into the liquid, however this will change the temperature of the wine (admittedly by a very small amount). If we put turnstiles into gates to count the number of people passing through, we change the speed at which people pass, and this can cause some people to change their route. If an app is downloaded onto a mobile phone to measure behavior it may change the performance and/or the battery life, which in turn may change the behavior being measured.

Researchers using observational data need to assess the extent to which there are observer and/or measurement effects. For example, are the outcomes of observed subjects the same as those for unobserved cases (after controlling for relevant sample differences)?

Survivorship Bias

Survivorship bias tends to occur when a researcher looks at a set of outcomes (for example, success or failure) and assumes that a) things that the successes share in common promote success, and b) things that are not shared by the success group are not promoting success. The sad failure of this logic was shown in a horrifying way by the Presidential Commission on the Space Shuttle Challenger Accident (Dalal, 2016). The problem that was found by the investigation was that O-ring seals used in the joints with the solid rocket booster had failed, due to cold weather (the temperature had fallen to 31°F).

As part of the investigation NASA showed a chart which highlighted all of the previous O-ring problems. This chart showed that they occurred at a wide range of temperatures, which had resulted in temperature being ruled out as a risk factor. NASA had estimated the risk of a shuttle failure at 1-in-100,000.

However, when the investigation looked at a chart that also showed the flights where no failures occurred the picture changed. When the temperature was above 65°F, there were 17 flights with no problems, and 3 flights with problems. In the four flights when the temperature was below 65°F there were problems in every case. Re-analysis of the data suggested that launching the shuttle at 31°F meant an approximate risk of shuttle failure of 13% (compared with the 1-in-100,000 figure based on survivorship bias).

A less serious, but equally illustrative case is the book *In Search of Excellence* by Tom Peters & Robert H. Waterman Jr. (1982). In preparation for the book the authors selected a group of successful companies and looked at what they had in common. The book reported these factors and concluded that these were the drivers of success. However, over the next few years many of these companies were not successful, casting doubt on Tom Peter's recommendation. Some of the factors these companies had in common were also shared with unsuccessful companies, but the research had not identified this.

To evaluate drivers from observational data it is necessary to take a good, representative sample of starts, not a selection of end points. The end points that are used should include both successes and failures (and all points between these two).

Not Explaining the Why

One of the most commonly heard criticisms of observational data, especially big data, is that it can't explain the why? In many cases this criticism is unfair. Travel data show that in cities like Amsterdam, when the weather is bad more people travel by public transport and fewer people walk and cycle. The models that have been created can forecast the shift from foot and cycles to public transport and the why is fairly obvious. Where a field is well understood, the patterns measured by data may be perfectly explicable in terms of the why.

However, in other cases the data do not explain the why. A good example of the limits of observational data and the why is given by Ben Wellington in his discovery of the location of the two highest grossing fire hydrants in New York, in terms of parking fines (2014). Using open data Wellington plotted all of the fines for all the fire hydrants in Manhattan (in New York it is illegal to park next to a fire hydrant). In many cases he was able to interpret the patterns, for example the upper East Side generated more fines because there were more hydrants. However, there were two almost adjacent hydrants in the lower East Side that were real outliers, generating about $55,000 a year in fines. Wellington visited the site of these two hydrants, took photos, and discovered that the issue was one of ambiguity. There was an informal but widely used bike lane between the road and the pavement, leading motorists to think it was OK to park, especially since parking spaces had been painted on the surface of the road. Wellington added a qualitative input to his data (observational qual data) and now understood the why as well as the what. The happy ending to this story is that after Wellington notified the City, they changed the road marking to make the situation clearer (and thereby reduced the number of cars parking illegally and the number of fines paid).

When working with observational data a researcher needs to assess if they are able to define the why without further research. If further research is needed, it might be another form of observational data (as in Wellington's observational qual), or it could be based on questions (e.g., surveys or focus groups).

Why Are Economists So Bad at Predicting Recessions?

In his book *The Signal and the Noise,* Nate Silver makes the observation that in 2008 most economists were still predicting that the US would not go into recession at a point when the retrospective data showed the US had already entered a recession. When trying to predict recessions, there are essentially two problems that economists face, the first is endogeneity and the second is a degrees of freedom problem.

Endogeneity refers to situations where the explanatory variables are influenced by other terms, in particular if there is a feedback loop between the dependent variable. Over the years, economists have discovered links between some element of the economy and a specific outcome. For example, William Phillips published a paper in 1958 showing an inverse relationship between rates of unemployment and increases in wage rates. However, as this information became assimilated into knowledge and actions of Governments, central banks, and financial markets it ceased to be true. A similar link between a pattern, a prediction, and the loss of the prediction's accuracy can be seen in the US at the moment in terms of the inverted yield curve. In the past when short term rates were higher than long term rates a recession ensued. Over the last couple of years this has not been the case, because predicted actions of the banks and Government have been factored into the behavior of the market.

In fields where the actors in a domain are aware of the models built from the data, there is a risk that their behavior will change because the model has been created. This is akin to the quote attributed to Albert Einstein: "No problem can be solved from the same level of consciousness that created it." To help identify this problem researchers should consider whether the people being modelled are likely to themselves use the results of the modelling.

The second problem, also highlighted by Nate Silver, is the degrees of freedom problem. Many of the organizations seeking to model the economy use hundreds of thousands of variables. However, there have been relatively few recessions. In terms of creating models that can predict a recession, the combination of just a few equations (i.e., each recession) and very, very large numbers of variables is a problem, a degrees of freedom problem. When the number of variables exceeds the number of equations by a modest amount, techniques such as hierarchical Bayes or Random Forests can help. However, when the number of variables massively exceeds the number of equations, the variables need to be simplified, as in the weather modelling example cited earlier.

CONFUSING INFLUENCE AND HOMOPHILY

Since the publication of books such as Malcolm Gladwell's *The Tipping Point* and Ed Keller & Jon Berry's *The Influentials,* there has been a whole industry promoting influencer marketing. However, in many cases the patterns observed in observational data do not represent patterns that will repeat themselves. One key alternative explanation to the patterns ascribed to influence is homophily. Homophily refers to the propensity of similar people to do similar things, captured by the phrase "birds of a feather flock together."

In 2007, a study was published looking at obesity and contagion and suggested that some people became obese because their friends were obese (Christakis & Fowler, 2007). The study concluded "Network phenomena appear to be relevant to the biologic and behavioral trait of obesity, and obesity appears to spread through social ties." This study was picked up by the general media and led to headlines such as "Are Your Friends Making You Fat?" (Thompson, 2009). The implication that was drawn was that people became overweight because they were influenced by their friends, i.e., that there was a causal link between having a friend who was obese and becoming obese yourself.

However, researchers such as Sinan Aral (2010) offer an alternative hypothesis, namely homophily. For example, do people become overweight because their friends are overweight, or do overweight people tend to go to the same place, do the same things, and become friends? In his research Aral demonstrates techniques for apportioning the contributions of influence and homophily. In many cases a particular phenomenon is the result of both effects.

Researchers should avoid automatically assuming that patterns they see in data are causal. Researchers should seek to quantify both influence and homophily when assessing patterns.

Observational Data and the Rear-View Mirror

One of the well-known challenges associated with observational data is the issue of it referring to past events. The question that confronts the user is the extent to which the future will behave like the past. In the short term, events tend to repeat themselves, tomorrow is often like yesterday, the next month is often similar to the same month last year. This challenge is associated with several of the issues raised earlier in this paper, for example correlation (which predicts the past), survivor bias (which assumes that the characteristics of the survivors are sufficient to cause the outcome), and influence (where the researcher needs to determine whether patterns are caused by influence or by some other factor).

Researchers working with observational data should seek to establish the extent to which conditions are expected to remain stable. At the very least the researchers should highlight assumptions. For example, models looking at car sales should highlight that regulations relating to a switch to electricity could invalidate the predictive power of models.

Remedies and Amelioration

The purpose of this paper is to highlight the challenges that observational data can present so that they can be addressed. Observational data presents many opportunities and these should be embraced, but they should be embraced judiciously. In each of the sections above along with the challenges a number of possible remedies or ameliorative measures are suggested. This section summarizes and expands these measures.

Controlled experiments are still seen as the gold standard. Where possible, experiments should be conducted, for example, by designing marketing campaigns in ways that allow the separate elements of the campaign to be evaluated.

Survey-enhanced models is a description of how traditional market research (both quantitative and qualitative) can be used to make models built on observational data more useful. Ben Wellington's visit to the site of the two fire hydrants mentioned earlier is an example of using qualitative research to enhance a model. When researchers use structural equation models to correctly attribute the effectiveness of marketing campaigns they typically include survey responses as one of the inputs, to show how changes in beliefs and awareness act as mediators.

Identifying the counterfactual is a key step in assessing causality. A control cell in an experiment is a counterfactual, matching people who have and haven't seen a social media campaign creates a counterfactual, and to a lesser extent previous years and expert predictions are counterfactuals.

Researchers should seek to minimize the number of independent (or predictor) variables and to maximize the number of independent observations. In classic, scientific, research the practice tends to be to only allow one variable to vary. In the real world of market research, it is often impossible to reduce the number of variables being modified to just one, but nevertheless, the researcher should seek to reduce the number of variables and to make them as uncorrelated as possible.

Alternative explanations should be sought for any pattern observed in a data set. Given a set of observations that suggest X causes Y, the researcher should be lead to postulate questions such as “How would we test whether Y is causing X?,” “How would we test that X and Y are caused by some third factor?,” “How can we check whether the relationship persists, i.e., that is not caused by chance?”

One exciting and relatively recent method of assessing causality in observational data is termed causal inference and that is the topic of the next section.

Causal Inference

Traditionally, statisticians have been very happy to report on associations and to accompany their measurements with statements about their accuracy and probability (for example 95% confidence that the number is 50% plus or minus 3%). However, these same statisticians, following in the tradition of Karl Pearson, have been reluctant (or completely unwilling) to make even probabilistic assessments about causality, unless that data were generated from controlled experiments. This reluctance has been challenged by the emergence of the field of causal inference. Causal inference has been developed by innovators from different disciplines, for example Daniel Rubin, Judea Pearl, and James Heckman. These innovators have challenged the traditional views of statisticians and treat causality as suitable for a probabilistic assessment.

In fields as diverse as economics, epidemiology, and the social sciences causal inference is being used and explored.

> “For example, in the technical program of the 2003 Joint Statistical Meeting in San Francisco, there were only 13 papers presented with the word “cause” or “causal” in their titles; the number of such papers exceeded 100 by the Boston meeting in 2014.” (Pearl et al., 2016)

There are at least two clear benefits from adopting the methods of causal inference. The first is that they provide a better way of describing the problem, and a better way of applying some of the traditional processes (for example, when to control for other variables and when not to). Judea Pearl is a great advocate of causal diagrams and a new algebra that helps frame the way the problem should be addressed. Daniel Rubin is an advocate of a “framework of potential outcomes,” which also obliges the researcher to consider the problem in a structured and systematic way (Rubin, 2011).

The second benefit, of course, is that in some cases the techniques are able to make a probabilistic estimate of the causal links.

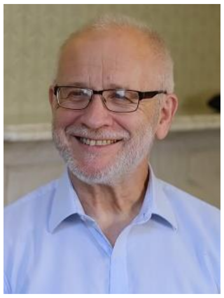
One challenge with causal inference at the moment is that there is not a settled view about which of the proposed techniques is best for which type of problem. There is a degree of tension between Pearl, Rubin, and Heckman (and much more between some of their acolytes). However, this is a field that is growing and one that holds great promise. With the growth in observational data and the desire to utilize observational data, causal inference could well be an idea whose time has arrived.

Conclusion

As the world becomes more digital, the amount of observational data is growing exponentially. This growth in data and the concerns about the ability of questions to generate valid and complete answers have led to an explosion in the use of observational data to answer questions and advise decision makers.

Observational data is a great resource for researchers, but there are challenges in using it that need to be recognized and dealt with. This paper lists a range of potential problems and, for each of them, steps that can be taken to tackle them.

The key point that needs to be made, is the point that is made by the advocates of causal inference, statistics needs to move on from being comfortable with association but shunning causality. Statisticians need to estimate, probabilistically, both the values of association and of causality.



Ray Poynter

References

1. Albright TD (2017) Why eyewitnesses fail. Proceedings of the National Academy of Sciences of the United States of America vol. 114,30.

2. Anderson C (2019) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. 23/6/2008, Wired, accessed 1 Nov 2019, https://www.wired.com/2008/06/pb-theory/

3. Aral S (2010) Social Contagion. Poptech 2010. Downloaded from http://opentranscripts.org/transcript/social-contagion/ on 6/11/2019

4. Calude C, Longo G (2017) The Deluge of Spurious Correlations in Big Data. Foundations of Science 2017, Volume 22, Issue 3.

5. Christensen H (2015) Banking on better forecasts: the new maths of weather prediction. Guardian 8/1/2015, accessed from https://www.theguardian.com/science/alexs-adventures-in-numberland/2015/jan/08/banking-forecasts-maths-weather-prediction-stochastic-processes on 6/11/2019

6. Dalal N (2016) The Space Shuttle Challenger Explosion and the O-ring. PriceEconomics. Accessed from https://priceonomics.com/the-space-shuttle-challenger-explosion-and-the-o/ on 6/11/2019

7. Damasio AR (1994). Descartes' error: Emotion, rationality and the human brain. New York: Putnam.

8. Christakis NA, Fowler JH (2007) The Spread of Obesity in a Large Social Network over 32 Years. The New England Journal of Medicine 2007 Volume 357.

9. Dubin, D (2011) Causal Inference Using Potential Outcomes. Journal of the American Statistical Association. 2005 Volume 100, Issue 469.

10. Gladwell, M (2000) The Tipping Point. Little, Brown and Company.

11. Gruschwitz D, Schönduwe R (2017) Collecting travel data using smartphone-based GPS-Tracking and web-based trip diary. ESRA 2017, Lisbon, Portugal.

12. Hassard JS (2012) Rethinking the Hawthorne Studies: The Western Electric research in its social, political and historical context. Sage Journals Volume: 65 issue: 11.

13. Kahneman D (2012). Thinking Fast and Slow (UK edition). London: Penguin.

14. Keller EB, Berry J (2003) The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy. Free Press.

15. Lee D, Pate RR, Lavie CJ, Sui X, Church TS, Blair SN (2014) Leisure-Time Running Reduces All-Cause and Cardiovascular Mortality Risk. Journal of the American College of Cardiology Volume 64, Issue 5, August 2014

16. Madsbjerg C (2017). The Power of the Humanities in the Age of the Algorithm. New York: Hachette Books.

17. McCambridge J, Witton J, Elbournec DR (2014) Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. Journal of Clinical Epidemiology, 2014 Vol 67.

18. O’Riordan M (2012) The Not-So-Long Run: Mortality Benefit of Running Less Than 20 Miles per Week. Medscape 6/6/19, downloaded from https://www.medscape.com/viewarticle/765209 6/11/19

19. Pearl J, Glymour M, Jewell N (2016) Causal Inference in Statistics. Wiley. Kindle Edition.

20. Pearson K (1911) The Grammar of Science (Third Edition). London. Adam & Charles Black.

21. Peters T, Waterman RH (1982) In Search of Excellence. HarperBusiness.

22. Phillips, W (1958) The Relation between Unemployment and the Rate of Change of Money Wage Rates in the United Kingdom, 1861–1957. Economica.

23. Poynter R (2019) Five Winds of Change, Sweeping Away Traditional Market Research. GreenBook Blog 24/09/19. Downloaded 6/11/19 https://greenbookblog.org/2019/09/24/five-winds-of-change-sweeping-away-traditional-market-research/

24. Silver, N (2012) The Signal and the Noise: The Art and Science of Prediction. Penguin UK.

25. Thompson, C (2009) Are Your Friends Making You Fat? The New York Times Magazine 10/9/2009 downloaded from https://www.nytimes.com/2009/09/13/magazine/13contagion-t.html on 6/11/2019

26. Watts DJ, Beck ED, Bienenstock EJ, Bowers J, Frank A, Grubesic A, et al. 2018. "Explanation, Prediction, and Causality: Three Sides of the Same Coin?." OSF Preprints. October 31. doi:10.31219/osf.io/u6vz5.

27. Weber T (2013) Response to "Run for your life . . . at a comfortable speed and not too far." Heart. 2013 Apr;99(8).

28. Wellington B (2014) Success: How NYC Open Data and Reddit Saved New Yorkers Over $55,000 a Year. I Quant NY, 2/6/2014. Downloaded from https://iquantny.tumblr.com/post/87573867759/success-how-nyc-open-data-and-reddit-saved-new on 6/11/2019

29. Four Vs, IBM Infographic, downloaded from https://www.ibmbigdatahub.com/infographic/four-vs-big-data, 30 October 2019.

30. ESOMAR (2019) Global Market Research Report 2019. ESOMAR Publications.

31. IPA 2015 The Expert Guide to Measuring Not Counting. IPA Publications.

32. FiveThirtyEight (2019) How Our MLB Predictions Work, accessed from https://fivethirtyeight.com/methodology/how-our-mlb-predictions-work/ 6/11/19

33. US Geological Survey, Can you predict earthquakes? Accessed from https://www.usgs.gov/faqs/can-you-predict-earthquakes?qt-news_science_products=0#qt-news_science_products on 4/11/2019