# PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

September 2016

# FOREWORD

These proceedings are a written report of the nineteenth Sawtooth Software Conference, held in Park City, Utah, September 27–30, 2016. One-hundred fifty 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 optimizing the design and craft of choice/conjoint analysis, surveying on mobile platforms, MaxDiff, market segmentation and classification, use of covariates in HB, and advances in market simulations.

The papers and discussant comments are in the words of the authors and very little copyediting was performed. At the end of each of the papers are photographs of the authors and co-authors. We appreciate their cooperation for these photos! It lends a personal touch and makes it easier for readers to recognize and greet 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 have helped this conference be such a success: Christopher Chapman, Keith Chrzan, Ken Deal, Joel Huber, and David Lyon.

Sawtooth Software
February, 2017

# Contents

# Summary of Findings

The nineteenth Sawtooth Software Conference was held in Park City, Utah, September 28–30, 2016. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2016 Sawtooth Software Conference Proceedings.

*** The Effects of Incentive Alignment, Realistic Images, Video Instructions, and Ceteris Paribus Instructions on Willingness to Pay and Price Equilibria** (Felix Eggers, University of Groningen, John R. Hauser, MIT, and Matthew Selove, Chapman University): The authors described how the decisions we make as researchers to craft our conjoint analysis surveys (with realistic images, video, incentive alignment, and appeals to the respondent to assume that the features are held constant if the attribute is not being shown) can affect price sensitivity, willingness to pay (WTP), and managerial decisions derived from the research. They conducted an experiment involving choice of smartwatches where respondents were assigned to different treatment cells that varied the elements of the CBC survey design. The use of high quality images affected the relative importances and increased the WTP for aesthetic elements of the smartwatch. The incentive alignment increased precision of the part-worths but interestingly enough also increased the derived WTP. The use of the video did not have much impact on the results other than to add time to the total survey length and perhaps increase respondent fatigue. However, it was pointed out that the aesthetic aspects of smartwatches perhaps did not need such an elaborate video to explain the features to respondents. The authors concluded by urging the audience to pay more attention and put more effort into the craft of conjoint surveys, particularly if conjoint analysis is used to simulate a market-based price.

* Honorable mention based on audience voting.

*** How Many Options? Behavioral Responses to Two versus Five Alternatives per Choice** (Martin Meissner, University of Southern Denmark/Monash University, Harmen Oppewal, Monash University, and Joel Huber, Duke University): Using eye-tracking technology, Martin and his co-authors conducted a detailed comparison of how respondents process pairs (2 concepts at a time) vs. quints (5 concepts at a time) for CBC studies. They found that respondents viewing pairs tended to pay attention to more attributes (out of the six attributes they studied) than the respondents viewing quints. The derived importances were more uniform for pairs compared to quints. Respondents viewing pairs tended to move their eyes from left to right across attribute rows between the concept pairs, whereas respondents evaluating quints more tended to move their eyes up and down within concepts. Respondents evaluating quints tended toward greater simplification (quickly discarding losing concepts) as key attributes were identified. Two somewhat unexpected findings from this research were that respondents rated the pairs as more difficult to complete than quints. Also, despite the lower statistical efficiency of pairs, utilities from pairs outperformed those from quints in predicting holdout triples. The authors stated that many consumer decisions (such as in FMCG) are better mimicked by showing a great deal of alternatives on the screen, but that there are many instances where conjoint pairs can be a good choice for reflecting the decision process and degree of attention to attributes commensurate with real world choices.

*Honorable mention based on audience voting.

**Findings of the 2016 Sawtooth Software CBC Modeling Prize Competition** (Bryan Orme, Sawtooth Software): Recently, Sawtooth Software held an open CBC modeling competition, attracting 15 teams to compete for a $5,000 prize. Naji Nassar of Marketing Intelligence and Research Services won the grand prize and the opportunity to present his model at this conference. Many different software systems, models, and model specifications were attempted, though there was a strong inclination to use HB models. The best model submitted by any one team gave a 5% reduction in error (RMSE of out-of-sample choices) compared to the standard default approach using Sawtooth Software's CBC/HB for utility estimation with randomized first choice (RFC) for choice simulation. This suggests that Sawtooth Software's default approach tends to work very well for this kind of CBC problem involving six attributes and vacation cruise choices. Interestingly enough, ensembles (via simple averaging of predictions) across the teams performed better than the best solution submitted by any one team. This held true for either in-sample hit rates or out-of-sample share predictions. Simulations on the level of the beta draws produced nearly identical results (though directionally slightly higher) than simulations using RFC on the point estimates. Although the use of ensembles could mean the difference between winning and losing a predictive modeling competition like this, Bryan wondered regarding the practicality of using ensembles in consulting practice to achieve the modest gains. There is a great deal of effort involved in generating the number of both diverse and high quality models needed for successful ensembles. Unless that could somehow be automated, it may be too much to undertake for typical client work.

**The Winning Choice Model: A Semi-Compensatory One** (Naji Nassar, MIReS): Naji described the approach he took to build his winning model for the 2016 Sawtooth Software CBC Modeling prize. First, he examined the means and distributions of choices using counting analysis. Using GAUSS, he developed an HB model that employed a Fuzzy Consideration Set approach—a semi-compensatory choice model that reduces IIA problems. Naji hypothesized that respondents might form consideration sets based on the cruise destination attribute and the overall cost (the budget) for the cruise vacation. He wrote a custom likelihood function to multiply the probability of selecting the destination x budget x the probability of selecting the alternative according to the logit rule based on all six attributes. Naji also employed logical utility constraints for some of the attributes.

**Using Bayes' Theorem to Adjust Simulated Preference Shares to Market Reality** (David Bakken, Foreseeable Futures Group): Conjoint researchers often find themselves looking to external effect adjustments to bring simulated shares of preference better in line with real market shares. David described previous efforts, including methods for accounting for product distribution and awareness. He then proposed an approach based on Bayes theorem that applies the known market share (or shares of preference from previous related conjoint research) as a prior. If applying market share as a prior, David's approach offers another way to leverage revealed and stated choice data. He demonstrated the result of the adjustments for real case studies involving home improvement/building products and health insurance. Via a series of market simulation results, he demonstrated that the Bayesian adjustment to shares has stable properties and face validity. David recommends the use of the Bayesian adjustment to shares if information about distribution and awareness or sales force effectiveness is unknown—and if you have solid information to apply as a prior, such as market share or shares of preference from multiple previous related studies.

**Mobile MaxDiff: What Are the Optimal Number of Attributes, Screens, and Level of Information Complexity?** (Michael Patterson, Radius Global Market Research, and Michael Smith, MFour): The authors revisited the question of how many items per set and number of sets for MaxDiff can be done effectively, but within the contemporary context of mobile interviewing. Mobile surveys often make up around 20% to 30% of completed surveys in developed countries. But, mobile is also associated with smaller screens and lower attention spans. Previous research at this conference has suggested that 4 or 5 items per screen is optimal and that interviewing on mobile is a viable option to fixed platform interviewing for MaxDiff. The authors devised a split-sample experiment to test if using 4 to 5 items per screen continues to be good advice for mobile MaxDiff surveys (sample drawn from Mfour's mobile panel which involved both smartphone and tablet users). They also manipulated the overall length of the MaxDiff survey and tested whether the length of the item descriptions (short or somewhat long) affected the quality of the results. They found that showing 7 instead of either 3 or 5 items per task resulted in greater abandonment rates and lower internal validity. Overall, the results were quite robust to the different treatments. Still, their research tends to confirm that showing 3 to 5 items works a bit better than 7 items per set. Not surprisingly, shorter questionnaires are associated with lower abandonment rates.

**Choice-Based Conjoint in a Mobile World—How Far Can We Go?** (Chris Moore and Christian Neuerburg, GfK): Given that completion of CBC surveys on mobile devices is now quite common, Chris and Christian conducted an ambitious split-sample experiment to test different ways of programming and setting up CBC surveys on mobile and to compare degradation in responses (if any) to CBC implementations on mobile devices. They programmed the surveys using different software platforms to test whether layouts that were responsive to mobile screens and orientation (respondents rotating their hand-held devices for portrait vs. landscape display) could affect the quality of the results. One of the surprises to the sponsors of the research was that mobile respondents tended to take surveys in quite a similar environment as for laptop/desktop responders: the vast majority at home and in a quiet/relaxed environment. However, the composition of mobile-completed surveys tended to skew younger and more female. Even the most demanding survey designs in the experiment (17 tasks, 10 attributes with not overly wordy text) still worked quite well in mobile, with very little self-reported difficulty or degradation in the survey experience. Pairs and triples tended to work best across all platforms and the authors recommended leaning toward asking more rather than fewer tasks whenever possible.

**Can Adaptive MaxDiff Provide Better Results than Standard MaxDiff?** (Howard Firestone, RTi Research): Adaptive forms of MaxDiff arrange items within sets based on respondents' previous answers. By referring to prior choices, better items can be compared against better items and worse items versus worse items in subsequent MaxDiff sets. Such a design, Howard argued, not only can be more discriminating at the individual level for measuring the items at the best and worst ends of the dimensions, but the interview can be shorter for the respondent and the process more compelling for the client. Howard conducted a split-sample test where respondents received one of four different versions of MaxDiff questionnaires: some non-adaptive standard approaches and some adaptive approaches. The adaptive MaxDiff cells led to greater discrimination among the items compared to the non-adaptive MaxDiff questionnaires. The predictions of holdout questions for the adaptive MaxDiff designs were comparable to the standard MaxDiff approach. Howard pointed out that the disadvantages of the adaptive MaxDiff approach include more programming expertise and data preparation time.

**Comparing Two Methods to Estimate Missing Maximum Difference Utilities** (Kelsey White and Paul Johnson, SSI): When working with a great deal of items in MaxDiff, some approaches (such as Express MaxDiff) lead to missing data (no information on certain items) for subsets of respondents. Paul and Kelsey designed an experiment that purposefully led to many missing evaluations of items at the individual level. They implemented a first series of MaxDiff questions (not covering all the 200 items) and then asked (in multiple rounds) if respondents wanted to continue answering more questions. Regarding analysis, HB estimation for Express MaxDiff imputes item scores at the individual level for missing items by taking draws from the upper level (the population means and covariances). Paul and Kelsey compared that approach to an EM algorithm informed only by the non-missing individual-level HB scores: they threw out the HB scores for the imputed items at the respondent level and then used EM to impute new scores. The subject of the study was a large MaxDiff study: 200 items representing statements that presidential candidate Donald Trump has made. At the aggregate analysis, the scores imputed using HB and EM were nearly identical, leading to an aggregate correlation of nearly 1.0. Segmenting respondents by stated party affiliation and comparing the scores also led to extremely similar conclusions, regardless of whether the imputation was done via HB or EM. Paul and Kelsey followed up with some of the same respondents a few months later to ask additional MaxDiff questions on items not seen by these respondents in the first wave of interviewing. They compared the imputed scores with the follow-up MaxDiff scores on the previously missing items. Hit rates at the individual level were essentially the same for HB and EM.

**The Researcher's Paradox: A Further Look at the Impact of Large-Scale Choice Exercises** (Mike Serpetti, Claire Gilbert, Gongos, and Megan Peitz, Sawtooth Software): MaxDiff has become extremely popular for placing multiple items on a common measurement scale and achieving enhanced discrimination compared to other item measurement approaches. However, clients often demand a great number of items for MaxDiff studies. The authors described a split-sample experiment in which respondents received one of six different large MaxDiff tasks. They compared Sparse MaxDiff, Express MaxDiff, and versions that included anchor questions or not. Some respondents saw as many as 60 MaxDiff screens (sets) whereas other respondents saw 30 or 18 sets. Sparse MaxDiff approaches tended to do a bit better than Express MaxDiff in predicting holdout respondents' choices. They found that for respondents receiving 60 sets, the first 30 sets performed nearly as well as using all 60 sets. This demonstrates sharply diminishing returns for asking respondents extremely long MaxDiff questionnaires. The authors also reported that showing respondents their scores in real time at the end of the survey tended to increase respondent satisfaction with the survey.

***Naïve Bayes Classifiers, or How to Classify via MaxDiff without Doing MaxDiff** (David Lyon, Aurora Market Modeling): Although the word naïve often connotes overly simplistic or bad, David pointed out that in statistics and for classification problems, naïve Bayes classifiers are very easy to implement and that they can work surprisingly well. As a first step, David showed how using simple cross-tab data on categorical survey questions and by applying Bayesian logic one can assign new respondents into any existing segment scheme. In doing so, the researcher assumes independence of the predictor variables and multiplies the likelihoods of observing the categories of the predictor variable given segment membership by the expected segment size (which serves as a Bayesian prior) to find the posterior likelihood of belonging to each group. Entirely aside from MaxDiff, naïve Bayes can be a general-purpose typing tool. Upon that logical foundation, David described how to incorporate MaxDiff questions into a typing tool as well. If the original segmentation scheme was based on MaxDiff, adding MaxDiff tasks to the typing tool helps accuracy considerably. With MaxDiff tasks, one extends the naïve Bayes classifiers by continuing to multiply across the likelihood that we would see respondents within each segment answer the MaxDiff questions in the manner that the new respondent taking the typing tool questionnaire did. David then described analytical and search approaches to find MaxDiff typing questionnaires that employ relatively few MaxDiff questions. He demonstrated using real case studies how successful MaxDiff typing tool questions (with many or few questions) can be for assigning new respondents to an existing segmentation scheme.

*Best paper award based on audience voting.

**Typing Tools in the Context of Choice Experiments** (Lech Komendant, IQS): Lech described different typing tool strategies for classifying new respondents based on an existing segmentation scheme (where that segmentation scheme could have been developed from MaxDiff, CBC, ACBC, or another preference measurement technique). He focused on three main strategies: 1) pairwise classifiers, 2) full rankings classifiers (multinomial regression based on ranks), 3) naïve Bayes tailored for MaxDiff, and 4) naïve Bayes tailored for MaxDiff plus an adaptive questioning component to potentially boost classification success. Lech used simulated respondent data (based on segments developed from real data and individual-level HB parameters) to test the different approaches. He concluded that all four approaches can be useful, but naïve Bayes and the rankings classifier were the best. Holding the number of typing tool tasks constant, including more items or profiles per task improves the results (but at the cost of longer completion time). The adaptive approach was not especially successful: it required a great deal more programming effort and led to only modest gains.

**Full-Flavoured HB: BYO Data in the Upper Model** (Jane Tang, Rosanna Mau, MARU/Matchbox, and Mona Foss, Bootstrap Analytics): Jane and her co-authors pointed out that BYO (build-your-own configurator) questions are somewhat common in product development research and are even a standard question type within Sawtooth Software's ACBC tool. Previous research on BYO questions has found that they can be useful as a training/education tool prior to conjoint questions, they encourage respondents to focus on each attribute, and they can potentially impact derived price sensitivity in the subsequent conjoint exercise. Rather than throw the BYO data away or encode the responses into the choice data, the authors included the BYO responses as covariates in the upper-level HB model. Covariates are most useful if they relate to attribute preferences and BYO questions on the conjoint attributes would seem by definition to fit that requirement. Jane and her co-authors showed how BYO questions as covariates affected the results for three real studies: one MaxDiff and two CBC studies. They found that the use of BYO as covariates adds nuanced, subtle, yet meaningful variation to the respondents' part-worth utilities—it captured greater heterogeneity and in their words brought out the "full flavour" of HB. Regarding whether using BYO questions as covariates improves the predictive validity of the models, they found only modest evidence of this and suggested it depends on the amount of heterogeneity in the data (disagreement across respondents). The more disagreement, the more opportunity there is for predictive gain by employing BYO questions as covariates. The use of BYO questions in the upper-level HB model provides a ready-made solution for generalization to future samples. The BYO questions themselves become the “golden” questions that can be quickly administered to the new respondents—allowing researchers to apply the HB model to the new sample without the conjoint exercise.

**Simulating from HB Upper Level Model** (Peter Kurz, TNS Infratest and Stefan Binner, BMS Marketing Research + Strategy): Over the last few years, a few leading researchers and academics have suggested that some conjoint analysis situations are better served by conducting market simulations using only the upper-level HB parameters (influenced by high quality covariates) and by actually ignoring the lower-level individual-level data. Peter and Stefan described how simulating from the upper level involves generating respondent agents whose characteristics are drawn from the respondents. They analyzed six CBC studies and compared the results of simulating based on a) respondent-level point estimates, b) the respondent-level draws, and c) simulating from the upper-level model (based on the population means and covariances). They reported mixed results regarding whether simulating from the upper level or the lower level provided better predictions of holdouts (either in-sample or out-of-sample). But, for projects involving product configuration (where the authors hypothesize that there is more uncertainty due to the large number of attributes and levels) the upper-level model performed better for out-of-sample holdout predictions. The authors concluded that when the conjoint study is complex and the sample size is relatively small, investment in obtaining good covariates for use in upper-level model simulations could be a valuable path.

**Mapping Attribute Non-Attendance** (Keith Chrzan, Sawtooth Software and Joseph White, MaritzCX): Respondents to conjoint surveys do not always pay attention to all the attributes, Keith and Joseph explained. They also pointed out that buyers in the real world also may not pay attention to all attributes. Numerous approaches have been proposed for measuring attribute non-attendance (the act of ignoring certain attributes), including stated measures, derived measures, and eye-tracking. Keith and Joseph used a threshold of importance approach (via a coefficient of variation hurdle) based on individual-level parameter estimation to infer the degree of non-attendance and then examined a number of commercial conjoint data sets regarding that measure. They found that attribute non-attendance was more prevalent in full-profile CBC and less prevalent in ACBC, best-worst conjoint (BW Case II), and partial-profile CBC. Keith and Joseph experimented with non-attendance indicators as covariates in HB, but saw little to no improvement in terms of model fit. Why worry about attribute non-attendance? The authors suggested that certain deliverables such as willingness to pay can be strongly affected by non-attendance. Also, researchers should think carefully regarding how much attribute non-attendance exists in the real-world buying decision and to choose a conjoint analysis approach that closely mimics and encourages similar choice behavior.

**Using Discrete Choice to Help Individualize Customer Lifetime Value** (Michael Smith, Michael Remington, and Michael Drago, The Modellers): Following recent papers and tutorials by academics such as Peter Fader, clients often prefer market segmentations that will help them focus their energy and advertising dollars on buyers who represent higher potential lifetime value. Michael and his co-authors explained that despite the appeal of these approaches, they typically focus on aggregate level models and information. They proposed a way to use CBC to compute customer lifetime value at the individual level. To illustrate their approach, the authors described a case study involving a professional sports team who wanted to incorporate customer lifetime value into a segmentation study. The discrete choice study explored respondent reactions to different motivations for their spending behavior (ticket prices, promotions, strength of schedule, preferred days of the week to attend games, etc.), allowing the researchers to forecast future spending behavior. The CBC task mimicked the look of a website where a person chooses a seat in a stadium and purchases a certain volume of tickets. The authors included other key pieces of information for targeting respondents in the survey to make the segmentation on lifetime value more actionable to the marketing department. Mike and his co-authors used the results to score a database of existing customers into the different segments. They also developed a typing tool for assigning new additions to the database into the segments.

**MTurk Survey Deception: Sources, Risks, and Remedies** (Kathryn Sharpe Wessling, Wharton School, Joel Huber, Duke University, and Oded Netzer, Columbia University): Kathryn and her co-authors related how common it is for academics to use Amazon's Mechanical Turk (MTurk), an online crowdsourcing labor market used for (among other tasks) completing social and market research surveys. Although the demographics of MTurk respondents are loosely representative of the general population, the age skews a bit younger. Even if an MTurk sample isn't perfectly representative, Kathryn explained that it provides a much cheaper, faster, and easier way for researchers to collect data compared to a professional panel company. But, cheating and deception can be a concern within MTurk, particularly when it comes to selecting subpopulations from the MTurk community. With multiple online communities providing tips for fellow "Turkers" to qualify for the highest paying surveys and hints for how to lie in order to get through the screener questions, and pass consistency traps Turkers can easily impersonate to maximize their payout. Kathryn and her co-authors developed a series of questionnaires to quantify the degree of deception and found some concerning results. Rather than deceivers just increasing the noise in survey responses, the authors showed that they systematically bias the responses to survey questions, including conjoint analysis. Moreover, deceivers over-report ownership or interest in any category that they think the researcher is screening for. They also tend to under-use the None category in CBC questions. Kathryn and her co-authors recommended that researchers who want to use MTurk develop their own panel of "Turkers," that they conduct their screening questions outside the primary survey, and that they monitor MTurk online communities during data collection to be on the lookout for collaborative cheating behavior. Kathryn recommended that professional panel companies, who may be losing survey business to the advent and popularity of MTurk, have the opportunity to innovate, by automating the panel quote process, reducing the speed to completion, and simplifying the backend payment process in order to compete with the MTurk platform. Given the issues with MTurk deception, professional panel companies can justify charging a premium, although the prices of professional panel companies are still relatively too high to compete with a MTurk sample.

**Process Tracing: A New Tool for Modeling Physician Treatment Algorithms** (Stephen Bell and Douglas Willson, A+A Bell Falla): Many research studies aim to understand how physicians make treatment decisions for specific patients facing certain health circumstances. Although conjoint analysis has been used for this purpose, Stephen and Douglas described a related but staged approach called process tracing (versions of which have been available to researchers since the 1970s, e.g., information boards). With process tracing, doctors are asked to make a decision regarding how to treat a patient once they feel they've been given enough information. Rather than give all the information upfront about the patient (and/or the treatment characteristics) they are shown a number of attributes that are initially blank (as if they were information cards that had been flipped over). The attributes are labeled (such as "age of patient," "patient prior history," etc.) but the level corresponding to that patient is only seen if the doctor clicks on the attribute to "turn the card over." After the doctor has turned over enough cards to gain enough information such that they are confident making a decision, the choice is made and the next task is shown. An ordered sequence of diagnostic steps (often following a tree-based hierarchy decision process) is common to doctors and encouraged in their medical training. Stephen and Douglas explained that this heuristic process seems quite natural for doctors to follow and is therefore realistic for understanding their decision making. Prior to constructing a process tracing survey, upfront qualitative work is usually required to understand the key attributes involved and the language the doctors use. Regarding the analysis of the process tracing survey data, multinomial logit (estimated via HB) may be used to model the heuristic decision process and build choice simulators similar to conjoint analysis. The importance scores tend to be quite similar between CBC and a simple reverse-ranking score for attributes chosen (uncovered) by the physician respondents.

**Let's Take None Seriously** (Ula Jones, Tomer J. Ozari, and Peter Kurz, TNS): Ula and her co-authors reviewed the various ways that have been proposed to elicit the None response, including standard, dual-response None (2-point and 5-point variations), and follow-up purchase intent questions. Although many researchers have begun to favor the dual-response None, the authors point out that this approach is not without its problems, one of which is that they argue that it just feels unnatural to respondents. Ula and her co-authors created a split-sample study to test four different ways of asking the None alternative: traditional, dual-response, reversed dual-response, and traditional None with follow-up product selection. With the reversed dual response, respondents are shown a CBC task with multiple product concepts and are first asked if they would purchase any of the products (yes/no). No matter whether they say yes or no, they are then asked which of the products they would be most likely to purchase. The traditional None with follow-up product selection only triggers a follow-up question if respondents pick the None alternative. In that case, they are shown the product concepts they just had seen and asked which of these they would be most likely to purchase. The authors found that respondents tended to like the traditional None a bit more than the other approaches (result not significant) and that the reversed dual response was preferred a bit more compared to the dual response None. The use of the None was highest in the reversed dual-response None format and least in the traditional None with follow-up.

*** The Art and Science of Nested Logit: Case Studies from Modeling Many SKUs** (Kevin Lattery, SKIM Group): Especially in FMCG studies involving many SKUs on the shelf, there are competitive (differential sourcing) effects that may not be captured well with logit models operating under the IIA (independence from irrelevant alternatives) assumption. Often the same brand with multiple SKUs will see its new (or improved) new offerings compete more heavily with its existing offerings. Oftentimes competition is enhanced among SKUs within the same package size. Although individual-level models under HB helps alleviate some of these concerns, the larger the models (the more parameters) and the more sparse the data, the greater the problems that remain with proper substitution and sourcing of volume. Kevin described how nested logit provides a framework for allowing the researcher to specify that certain groups of product alternatives (nests) compete more heavily within the nests than between the nests. Furthermore, nested logit provides a way to estimate that degree of competition (correlation) among alternatives within the nests. How to choose the nests is a challenge and Kevin demonstrated a counting approach (computing the overlap between pairs of SKUs) followed by a hierarchical clustering to help the researcher develop hypotheses regarding appropriate nesting structures. Kevin recommended testing different nesting structures with an aggregate nested logit model, then pruning nests that don't seem to be justified. For final modeling, Kevin recommends using nested forms of latent class or HB (he suggested R packages for this), though there are challenges involved in model specification, convergence, and time to run the models.

* Honorable mention based on audience voting.

**Mining and Organizing User-Generated Content to Identify Attributes and Attribute Levels** (Artem Timoshenko and John R. Hauser, MIT Sloan School of Management): Although conjoint analysis has been successful in many applications, the challenge always facing the researcher is how to select the right attributes and levels for the study. Artem and John reviewed common approaches, including expert opinion, competitive research on the internet, reviewing user-generated content (UGC) in the form of customer reviews, or needs-based approaches which translate customer needs into attributes. The hurdle with UGC is the shear amount of information which can sometimes involve gigabytes of typed language. To identify customer needs from UGC, the authors designed a machine learning approach based on Convolutional Neural Networks (CNN) and dense sentence representations to identify informative content and sample a diverse and extremely reduced subset of the content for humans to then review. The authors compared how many unique customer needs could be identified if humans randomly sampled sentences to process and code manually as opposed to letting machine learning methods screen down the phrases for further human review. They found that the machine learning-aided process was substantially more efficient and consistently led to discovery of more unique customer needs. The authors concluded by suggesting that future improvements may involve completely automating all aspects of the machine-learning algorithm so that no human intervention is needed.

**What a Difference Design Makes** (Karen Buros, Radius GMR, and Jeremy Christman, Procter & Gamble): Karen and Jeremy described the challenges and successes they encountered when using a menu-based choice approach to understand how buyers purchase products within a personal care category. The product category involved three compatible product types (cleansing, finishing, and remedial) which consumers regularly use on all or single occasions, three major brands, and many overlapping benefits. Respondents were shown a subset (22) of the 86 different products in each of 12 choice tasks and could multi-select as many or as few as they wanted to choose within each task. The authors initially built 86 separate binary-logit HB models with ASCs and all significant cross-effects, but faced some challenges getting predictions from the resulting market simulator that could reasonably match raw count data. Cleaning the data of respondents who answered in clearly illogical ways, pruning the model to only involve the most significant cross-effects, and constraining the signs of those cross effects all seemed to help. But, the biggest breakthrough in bringing the market simulator in line with the raw counts data was to recognize how much extrapolation was involved if building a simulator that assumed that all 86 items were being shown at once, when the data generation process involved respondents only ever seeing 22 items at a time. Subsequent refinements involved collapsing the design space from 86 total items into smaller factors to test the specific hypotheses of the research. Karen and Jeremy emphasized the importance of taking enough time to undertake the necessary steps to do the job right for complex MBC studies like these.

**Explaining Preference Heterogeneity with Mixed Membership Modeling** (Marc R. Dotson, Brigham Young University, Joachim Büschken, Catholic University of Eichstatt-Ingolstadt, and Greg M. Allenby, Ohio State University): Marc and his co-authors acknowledged that finding covariates that are predictive of part-worths is challenging, but evidenced that more could be done to include more covariates in HB modeling of choice data via dimension reduction in the covariates by accounting for co-occurrences. They described a dataset where respondents had completed a pick any/j questions regarding the benefits that respondents believe robotic vacuums could bring them, followed by a standard CBC questionnaire on robotic vacuums. Rather than use all 18 benefit items as binary covariates in the model, they utilized a grade of membership model (GoM). The GoM allows data reduction into various latent factors and the computation of a membership vector (loadings on the latent dimensions) for each respondent. What makes the data GoM reduction technique different from other approaches is that it uncovers extreme preference profiles (similar to archetype analysis) rather than summary profiles more representative of the means of latent dimensions. Marc and his co-authors found that the use of the GoM membership vector as covariates in HB estimation could improve the in-sample fit of the data as well as the out-of-sample hit rate for holdouts.

# THE EFFECTS OF INCENTIVE ALIGNMENT, REALISTIC IMAGES, VIDEO INSTRUCTIONS, AND CETERIS PARIBUS INSTRUCTIONS ON WILLINGNESS TO PAY AND PRICE EQUILIBRIA

**FELIX EGGERS**[1]
*UNIVERSITY OF GRONINGEN*
**JOHN R. HAUSER**[2]
*MIT*
**MATTHEW SELOVE**[3]
*CHAPMAN UNIVERSITY*

## ABSTRACT

We describe how craft in conjoint analysis surveys (realistic images, incentive alignment, training videos, and ceteris paribus instructions) affects both accuracy (relative part-worths) and precision (scale of the part-worths). Accuracy and precision, in turn, affect estimations of willingness to pay (WTP) and predictions of market-equilibrium prices and profits for various "what-if" scenarios. Managerial recommendations, which attribute levels to include in a product and how to price a product, vary dramatically depending upon the craft of the study. When used in litigation, craft also affects the estimated value of copyrights and patents. To demonstrate the effect of craft, we conducted an experiment (smartwatch application) in which we systematically varied different drivers of craft. The use of realistic images increased accuracy and precision. Incentive alignment increased precision, but not accuracy. Neither training videos, nor ceteris paribus instructions had a positive effect. In fact, training videos reduced precision substantially because the wear-out effect (for our data) overwhelmed the training effect. The effect of craft on accuracy and precision had dramatic effects on estimations of WTP and equilibrium prices and profits. Managerial recommendations depended critically on craft as well as whether precision was based on the estimation data or adjusted for external validity. Craft matters! Defaults are not sufficient and could lead to incorrect recommendations and valuations.

*This paper summarizes results from Eggers, Hauser and Selove (2016). All copyrights remain with the original paper, which provides much greater detail. Non-exclusive permission is given to Sawtooth Software to publish this paper.*

## MOTIVATION

Modern conjoint analysis has been successful in the sense that it is now relatively easy to implement a conjoint analysis study. For example, students using Sawtooth Software's Discover package can create conjoint analysis designs and questionnaires within minutes with simple-to-understand point-and-click methods. Although pictures and animations are feasible, the default in most software packages is that profiles are described using plain text. Furthermore, advanced

[1] Felix Eggers; University of Groningen, Faculty of Economics and Business
[2] John R. Hauser; MIT Sloan School of Management, Massachusetts Institute of Technology
[3] Matthew Selove; George L. Argyros School of Business and Economics, Chapman University

methods such as incentive alignment and instructional videos are difficult and expensive to implement. Not surprisingly, many conjoint studies rely on defaults. But does it matter? Does "craft" matter?

Every day, both academics and practitioners make cost vs. benefits decisions about how to implement a conjoint analysis study. Higher craft, e.g., more realistic pictures or animations, are often expensive and time-consuming. We would like to know whether the additional cost is justified. For example, are managerial recommendations sensitive to craft decisions such as the selection of images for products and attributes, the use of incentive alignment, the use of video instructions, enhanced instructions that "all else is equal," the use of dual-response formats, the number of alternatives in a choice set, the number of choice sets, the inclusion of attributes that describe the product in question but are not of managerial interest, etc.?

The question of craft is extremely important. Not only are there roughly 14,000 conjoint analysis applications per year (Orme 2009, p. 127), but conjoint analysis is now being used to value copyrights and patents, often resulting in judgments in the hundreds of millions of dollars (Cameron, Cragg, and McFadden 2013). Through theory and empirical examples, we demonstrate that craft does matter. Craft affects pricing decisions, strategic decisions on which attributes to include in a new product, predictions of market response, predictions of profits due to managerial actions, and the valuations attached to copyrights and patents.

## Craft Affects Both Accuracy and Precision

For the purposes of this paper, we distinguish two aspects of a conjoint analysis study that might be affected by craft: accuracy and precision. We call a conjoint analysis study *accurate* if it estimates the correct *relative* part-worths. We call a conjoint analysis study *precise* if it estimates the correct scale, that is, the correct *absolute* magnitudes of the part-worths. Precision measures the signal-to-noise ratio, because it compares that which is explained by the attributes in the conjoint design to that which remains noise (the error term).

To define accuracy and precision mathematically, consider the standard logit model that is used in most choice-based conjoint analyses. For ease of exposition, we write the equation for binary attributes and for dummy-variable coding. The concepts apply to effects coding and to multi-level attributes. Indeed, our empirical example includes a multi-level attribute.

Let $u_{ij}$ be consumer $i$'s utility for product profile $j$. Let $\delta_{jk}$ be a binary indicator for the $k^{th}$ attribute such that $\delta_{jk} = 1$ if attribute $k$ is at its "higher" level for profile $j$ and $\delta_{jk} = 0$ if attribute $k$ is at its "lower" level for profile $j$. Although we say "higher" and "lower," we do not require that the higher level be preferred to the lower level, or that the ordering of levels is the same across consumers. For example, the "higher" level might be a silver-colored product and the "lower" level might be a gold-colored product. Let $p_j$ be the price associated with the $j^{th}$ product profile.

To fully specify the utility function, we define $\beta'_{ik}$ as consumer $i$'s (raw) part-worth for attribute $k$ ("higher" vs. "lower" level), $\eta_i$ as the weight for price, and $\epsilon_{ij}$ as an i.i.d. Gumbel-distributed error term. Assume there are $K$ attributes. Utility for the $j^{th}$ product profile is specified as:

$$u_{ij} = \sum_{k=1}^{K} \delta_{jk}\beta'_{ik} - \eta_i p_j + \epsilon_{ij} \tag{1}$$

The logit model predicts the probability, $P_{ij}$, that consumer $i$ chooses the $j^{th}$ product profile for a choice set consisting of $J$ product profiles. In this equation, we let $u'_{io}$ denote the utility of the no-choice option.

$$P_{ij} = \frac{e^{\sum_{k=1}^{K} \delta_{jk}\beta'_{ik} - \eta_i p_j}}{\sum_{\ell=1}^{J} e^{\sum_{k=1}^{K} \delta_{\ell k}\beta'_{ik} - \eta_i p_\ell} + e^{u'_{io}}} \tag{2}$$

Following McFadden (2014), we rescale utility to include a scale factor, $\gamma_i$, such that the relative weight on price is 1.0. In this formulation, as interpreted by McFadden (2014), the $\beta_{ik}$'s are the amounts that respondent $i$ is willing to pay (WTP) for moving attribute $k$ from its "lower" level to its "higher" level. (If the lower level is preferred, the WTP is negative.) Note that the WTP does not depend upon $\gamma_i$.

$$P_{ij} = \frac{e^{\gamma_i(\sum_{k=1}^{K} \delta_{jk}\beta_{ik} - p_j)}}{\sum_{\ell=1}^{J} e^{\gamma_i(\sum_{k=1}^{K} \delta_{\ell k}\beta_{ik} - p_\ell)} + e^{\gamma_i u_{io}}} \tag{3}$$

We call $\gamma_i$ the precision (for consumer i). In the conjoint analysis literature, $\gamma_i$ is sometimes called the "scale factor." The basic concept is that if $\gamma_i$ is large, then the standard deviation of the error term is small compared to the magnitude of the part-worths. A small *relative* error term means that the CBC logit model predicts more precisely the consumer's choices. (We say "relative" because, in most logit specifications, there are $K$ parameters for $K$ part-worths. Thus the standard deviation of the error term is not identified independently of the part-worths—only the relative magnitude of the error term is identified.)

We illustrate accuracy and precision with Table 1. Consider a conjoint design with three binary attributes and price. Suppose that the true raw part-worths represent how consumers actually behave in the marketplace when making choices among products described by these three attributes and price. (The error term includes all unmodeled effects including attributes that are not accounted for and any inherent uncertainty in consumer choice.) These part-worths are shown in the first column of data in Table 1.

**Table 1. Illustration of Precision and Accuracy**

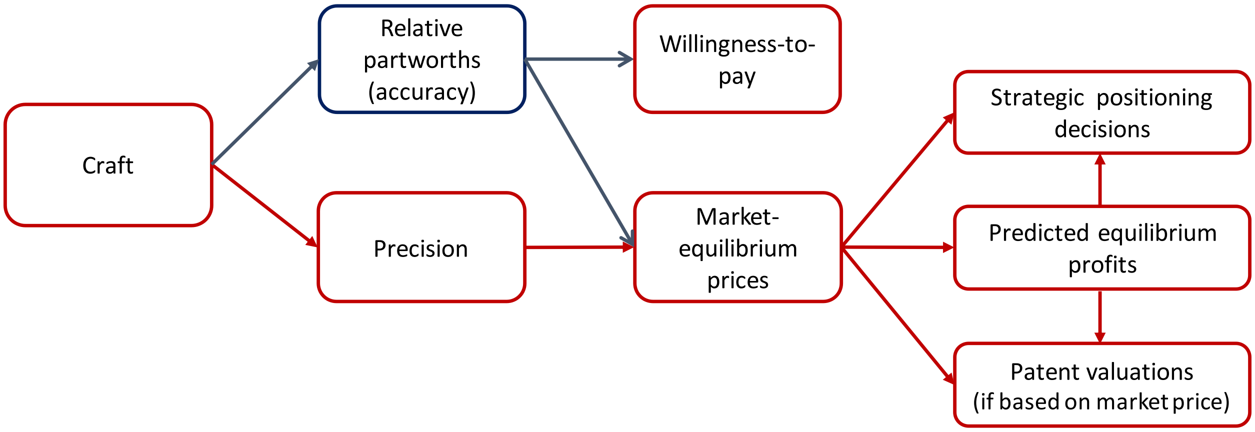
| | True (raw) part-worths | Lower accuracy | Lower precision | Higher precision |
|---|---|---|---|---|
| **Attribute 1** | 1.0 | 2.0 | 0.50 | 2.0 |
| **Attribute 2** | 2.0 | 1.0 | 1.00 | 4.0 |
| **Attribute 3** | 0.5 | 0.5 | 0.25 | 1.0 |
| **Price** | 1.0 | 1.0 | 0.50 | 2.0 |

Suppose that we estimate a model that has roughly the same magnitude of part-worths, but it switches the importances of attributes 1 and 2. We say that such a model has lower accuracy (2nd column of data). Suppose we estimate a model that gets all *relative* part-worths correct, but has a lower scale. We say that such a model is accurate but less precise (3rd column of data). Finally, the last column of data illustrates a model that is accurate but appears (to the analyst) more precise than truth.

Equation 3 is a useful theoretical definition of precision, but, in a population of consumers we might prefer a definition of precision that takes into account the fact that both the relative part-worths ($\beta_{ik}$'s) and the precisions ($\gamma_i$'s) vary over respondents. The ability to model such variation is an important advantage of advanced models such as hierarchical Bayes or empirical Bayes. When part-worths vary or when accuracy differs between studies, a better empirical measure of the precision is the average of the respondents' sum of absolute attribute importances (Arora and Huber 2001). We use that measure in our empirical comparisons throughout this paper.

Having defined accuracy (relative part-worths) and precision (scale factor), we now hypothesize that craft affects both and we hypothesize that both accuracy and precision affect managerial recommendations. We illustrate these hypotheses in Figure 1.

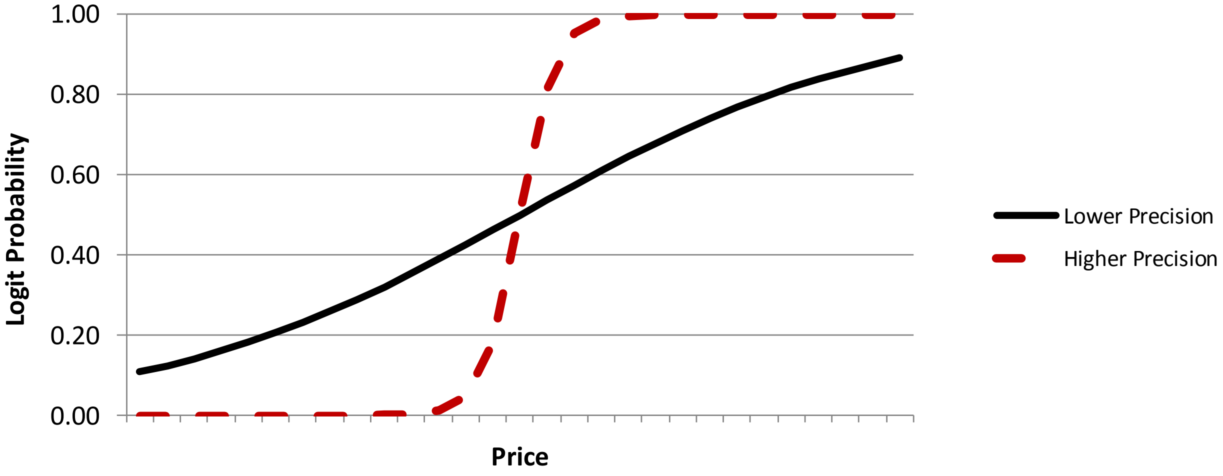
**Figure 1. Hypotheses: Craft Affects Managerial Recommendations**


Craft
Relative partworths (accuracy)
Precision
Willingness-to-pay
Market-equilibrium prices
Strategic positioning decisions
Predicted equilibrium profits
Patent valuations (if based on market price)

## Precision Affects Price Sensitivity, Which, in Turn, Affects Predicted Market-Equilibrium Prices and the Strategic Selection of Attributes for Products

When true precision is higher, consumer choices are more sensitive to changes in attributes or prices. We illustrate this phenomena in Figure 2 where we use Excel to plot the logit probabilities, $P_{ij}$, as a function of prices ($p_j$) for lower precision ($\gamma^{lower}$) and for higher precision ($\gamma^{higher}$). When precision is lower, as in the solid line in Figure 2, choice probabilities change more slowly as price changes. The curve is much flatter, almost a straight line. However, when precision is higher, as in the dotted line in Figure 2, choice probabilities change more quickly as price changes. The curve is much steeper. For sufficiently high precision ($\gamma_i \rightarrow \infty$), the logit model acts like a first-choice model (a step function).

**Figure 2. Higher Precision Means Greater Price Sensitivity**



## Willingness to Pay

WTP clearly depends upon accuracy. This is clear in McFadden's (2014) formulation because consumer $i$'s WTP for attribute $k$ is the relative part-worth ($\beta_{ik}$). If the $\beta_{ik}$'s are incorrect, then the estimate of WTP will be incorrect. While WTP is not market price (Orme 2009, p. 87), WTP is valuable in its own right. For example, when Polaroid launched the iZone camera it was able to determine that consumers would pay, on average, close to $10 for interchangeable camera covers—a feature that cost but pennies to produce. On the other hand, it learned that consumers were unwilling, on average, to pay anywhere near the cost necessary for the iZone camera to produce higher-resolution photographs. (The iZone camera was targeted to kids and produced postage-stamp-sized instant pictures.) Polaroid included interchangeable camera covers, but not higher-resolution capability (McArdle 2000). Similarly, when valuing patents, the marketing expert is often asked to provide WTP estimates to "damages" experts who combine secondary data with WTP to arrive at valuations (McFadden 2014; Mintz 2012).

## Market Equilibrium-Price

After a conjoint analysis model is estimated, the relative part-worths and precisions can be used in a choice simulator. Choice simulators predict how aggregates of consumers (the market and/or market segments) react to changes in attributes or price. Allenby *et al.* (2014) propose that conjoint analysis market simulations be used to compute Nash equilibrium prices. They further propose that the courts rely on the marketing expert to be both a marketing expert and a damages expert by estimating the change in Nash equilibrium prices due to a patent. They propose that the output of the conjoint analysis simulator for a product with the patented feature be compared to the output of the conjoint analysis simulator for a product without the patented feature.

The same methods, as those proposed by Allenby *et al.*, can be used to predict how the market will react to the introduction of a new product or to a change in a product's attribute levels. If costs are known, the simulator can predict profits for the new product or for a change in a product's attribute levels. The analysis could be extended to situations where competitors are hypothesized to respond to a new product by changing their attribute levels. For example, if BMW introduces adaptive cruise control, we might expect Audi, Mercedes, and Lexus to respond by introducing their own versions of adaptive cruise control. (By the way, this has already happened.)

As illustrated in Figure 2 (for price), predictions of consumer price response depend upon precision. This sensitivity to precision is particularly critical if the simulator is used to predict equilibrium prices. For example, Hauser, Eggers, and Selove (2016) illustrate the sensitivity of equilibrium prices to precision with a simple two-segment model in which the relative part-worths and precisions are homogeneous within segment (but not between segments). They then vary precision and compute the Nash equilibrium prices. These prices are shown in Table 2. Notice that equilibrium prices vary dramatically over the range of precisions that we might expect in empirical conjoint analysis studies.

**Table 2. Precision Affects Equilibrium Prices**

| Precision ($\gamma$) | Predicted Equilibrium Price in Differentiated Market (in currency units) |
|---|---|
| **0.5** | 2.59 |
| **1.0** | 1.42 |
| **2.0** | 0.92 |
| **3.0** | 0.82 |
| **4.0** | 0.79 |
| **5.0** | 0.78 |

In Table 2, the equilibrium prices vary from under 1.0 currency unit to over 2.5 currency units. This could have a dramatic effect on whether or not a product is launched or in the valuation of a copyright or patent. In a typical patent case, such a difference in equilibrium prices (with and without a patented feature) could mean a difference in valuation in the hundreds of millions, or even billions, of dollars. For example, Apple has sold over 800 million iPhones. If the difference in price due to a patented feature swung from $10 to $25 due to precision, that would imply a difference in valuation of $12 billion. (These prices are the equilibrium prices, not the differences in equilibrium prices. Estimating differences requires additional simulations, but the point is made. Prices depend dramatically on precision.)

If craft affects precision, then clearly craft matters for predicting price and profits, whether it be for a newly designed product, a change in a product's attribute levels, or due to a patented or copyrighted feature. It is, of course, obvious that accuracy affects WTP and hence also affects predicted equilibrium prices and profits.

## Strategic Recommendations for the Attribute Levels of a Product

Suppose that more consumers prefer a silver-colored watch to a gold-colored watch than vice versa. An innovator of smartwatches, facing no competition (and limited to one color) might introduce a silver-colored smartwatch. Even if the innovator is anticipating a competitor, the innovator might try to preempt competition and "position" its product as a silver-colored smartwatch. In practice, such strategic positioning decisions are usually made on product "positions" that are difficult to match. For example, Brita filters preempted competition by positioning themselves as the best-tasting water filters. It's competitor, P&G's PUR water filter, differentiated the market by positioning itself as healthy. The market was also differentiated on attributes, with Brita dominating pitcher filters and PUR dominating faucet filters.

A follower now has a choice. If the follower ignores the fact that the innovator is marketing a silver-colored smartwatch, the follower might be tempted to introduce a silver-colored smartwatch. For example, the follower might use a conjoint analysis simulator without taking into account that the innovator will respond to the follower's launch. For example, the innovator might lower its price to combat the market entry. A more sophisticated follower might decide to differentiate the market and introduce a gold-colored smartwatch. The sophisticated follower hopes that it will sell to the gold-color-preference segment leaving the silver-color-preference segment to the innovator, thereby reducing price competition.

The decision depends on the size of the two segments, on the market response to color, and on the market response to price. If the market is not very responsive to either price or color, differentiation will have little effect on reducing price competition. The follower might be best advised to target the largest segment and introduce a silver-colored smartwatch. On the other hand, if the market is very responsive to price or color, differentiation will reduce price competition. The follower might be best advised to introduce a gold-colored smartwatch.

Market response to both attributes and price depends upon precision. Hauser, Eggers, and Selove (2016) prove formally that lower precision implies an undifferentiated market, while higher precision implies a differentiated market. They also demonstrate that the intuition based on the formal analysis applies when heterogeneity is continuous as is modeled in standard conjoint analysis estimation. (While this result has the flavor of standard analyses of differentiation, the effect is due to precision *not* to heterogeneity in consumer preferences.)

Put another way, the strategic choice of which level to include in a product depends upon precision, even if the relative part-worths are perfectly accurate. Of course, accuracy also affects the relative part-worths and, hence, the choice of attribute levels for the firm's product. Craft matters for the strategic choice of product attributes!

## TRUE PRECISION VERSUS THE ESTIMATED PRECISION UPON WHICH THE FIRM RELIES

Before we examine empirically four elements of craft (drivers of precision), we pause to discuss interpretations of precision. Typically, in conjoint analysis applications, analysts estimate a logit model and report the part-worths. The absolute part-worths might be used in a simulation, as is common in academic studies. Alternatively, the relative part-worths might be used, combined with an analyst-chosen error magnitude, as in randomized first-choice simulations. In either case, there is a precision ($\gamma$) associated with the simulation. When this precision is based on the estimated logit model, that is, the choice sets used to estimate the logit model, we might call it estimated precision, $\gamma^{estimation}$.

Typically, analysts report *fit* statistics such as hit rate, the percent of uncertainty explained ($U^2$), Kullback-Leibler convergence, root likelihood, AIC, BIC, or DIC. Sophisticated analysts also report these statistics for holdout choice sets. Experienced analysts know that the holdout statistics are never as good as the fit statistics. The precision reported based on the fit data may not be the precision that applies to the holdout data. It is a simple matter to re-estimate precision in a one-variable logit model. The dependent variable in this logit model is choices in the holdout sets and the explanatory variable is utility as created by the *relative* part-worths. We might call this data-based precision, $\gamma^{internally\ adjusted}$. We use the word "internally" because holdout choice sets are really a test of internal validity. For high-craft studies, we might expect that internal validity is a good indicator of external validity, but that is worth testing. For one test, see Toubia *et al.* (2003).

Although rare, some analysts go a step further and test a form of external validity. For example, the analyst might create an incentive-aligned marketplace and ask respondents to make choices in that marketplace after a delay of a few weeks. The closer the induced marketplace is to the real marketplace, the closer the test is to a test of external validity. We adjust precision for external validity with the same type of one-variable logit. The only difference is that the dependent variable is now the choice in the induced marketplace. We might call this externally-adjusted precision, $\gamma^{externally\ adjusted}$.

We hypothesize that $\gamma^{internally\ adjusted}$ and $\gamma^{externally\ adjusted}$ depend upon craft. That is, we expect that higher craft leads the analyst to estimate models that fit the holdout data better and fit any induced marketplace data better. We expect that the precision from the highest craft study, adjusted to the induced marketplace may be our best estimate of true precision, $\gamma^{true}$. Estimation precision, $\gamma^{estimation}$, may or may not depend upon craft. If consumers are consistent in their answers to lower-craft questions, $\gamma^{estimation}$ might be high even if we cannot predict holdout choices or choices in an induced marketplace.

In this paper, we report $\gamma^{estimation}$, because this is the most common precision that is reported (when precision is reported). We also report $\gamma^{externally\ adjusted}$ based on an incentive-aligned induced marketplace with twelve smartwatches and an outside option that takes place three weeks after the estimation (and holdout data) are collected. (One in 500 respondents received the smartwatch they chose in the induced market, plus any change from $500.) We take on faith that $\gamma^{externally\ adjusted}$ is closer to $\gamma^{true}$ than is $\gamma^{estimation}$ or $\gamma^{internally\ adjusted}$.

## An Empirical Study to Test Drivers of Precision and Accuracy

We chose to test four drivers of precision and accuracy: (1) relative realism of the images used to describe attributes and profiles, (2) incentive alignment, (3) videos that train respondents about attributes and the choice tasks, and (4) instructions that all attributes, not displayed explicitly in the product profiles, are to be considered common to all profiles in the choice set (ceteris paribus instructions). These four drivers are cited often in the literature and in challenges to the use of CBC to value patents and copyrights. If we can show that any of these elements of craft affect managerial recommendations, then we hope to encourage other researchers to examine additional elements of craft.

The context of the empirical test is smartwatches. This category is sufficiently complex to make the test relevant, but not so complex as to make an initial test infeasible. We focused on just three attributes and price: case color (silver or gold), watch face (round or rectangular), watch band (black leather, brown leather, or matching metal color), and price ($299 to $449). Naturally, any missing features affect the error term, and hence precision, thus, $\gamma^{true}$ remains finite. The effect of non-modeled attributes should be constant across all but the ceteris paribus manipulation. In the ceteris paribus instructions, we inform consumers that brand, operating system, and technical features remain constant across all profiles. (Consumers are asked about their brand choice and all profiles are about that brand.) Because brand (and operating system) are important features of smartwatches, we hypothesized that ceteris paribus instructions (versus lack thereof) would affect precision. Holding brand and operating system constant means the ceteris paribus manipulation provides a strong test for the effect of non-modeled attributes in the smartwatch category.

We wanted to maintain high quality on all aspects of the study that were not experimentally varied. Such recommendations are made in Allenby *et al.* (2014). We recruited a US sample from a professional panel. The panel, Peanut Labs, maintains 15 million prescreened respondents. Peanut Labs is a member of the ARF, CASRO, ESOMAR, and the MRA and has won many awards for high quality. We targeted respondents who expressed interest in the category, were aged 20–69, and agreed to informed consent as required by our internal review boards.

We followed standard survey design principles (Schaeffer and Presser 2003). All questions, attributes, and choice tasks were pretested carefully (66 respondents using the same sampling criteria). We tested for and found no demand artifacts. We used hierarchical Bayes logit-based estimation (Sawtooth Software 2009). We used sixteen choice sets for estimation (and two for holdout validation) with three profiles per choice set. (Three profiles is, by far, the most common in applications [Meissner, Oppewal, and Huber 2016].) We used a dual-response format for the outside option (Brazell *et al.* 2006; Wlömert and Eggers 2016).

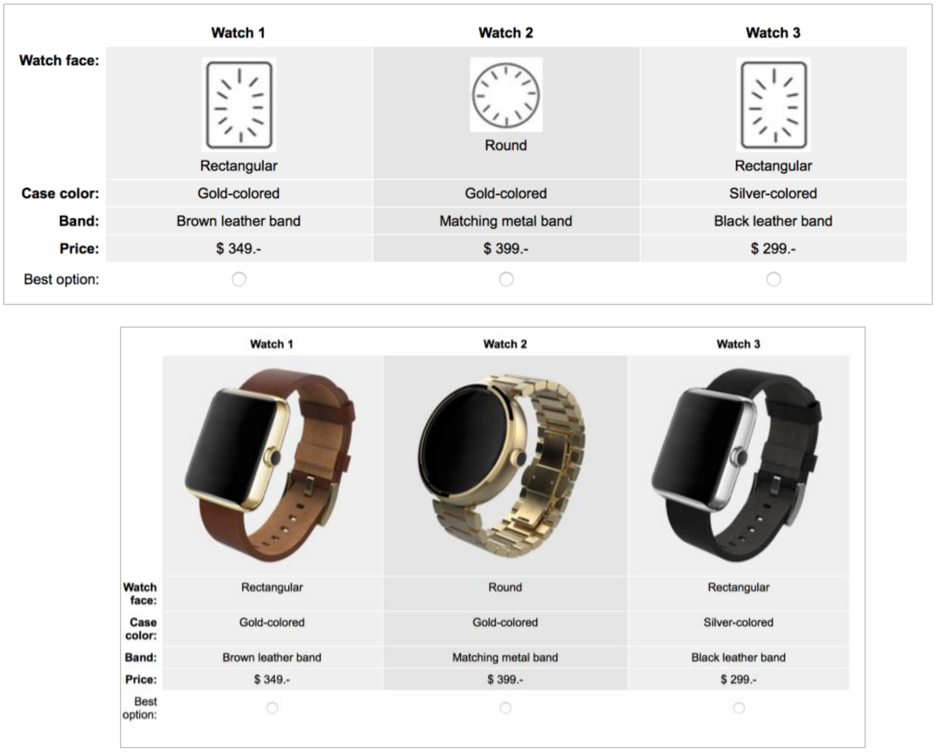
We varied the four elements of craft in an orthogonal half replicate of a $2^4$ design. Respondents were assigned randomly to each experimental cell of the half replicate, with roughly an equal number of respondents in each cell. Three weeks after the conjoint analysis studies were completed, we recontacted respondents and asked them to choose a smartwatch from a (full factorial) market of twelve smartwatches in a dual-response format. We assigned prices to the twelve smartwatches randomly and confirmed, based on the pretest data, that none of the options was dominating or dominated. The choice was incentive-aligned and presented the most realistic images of smartwatches.

The four elements of craft were:

### Realism of the Stimuli

For the (hypothesized) higher level of craft we used realistic images of smartwatches in which the attributes were embedded in the images and listed in easy-to-read text below the images. The images included animations that allowed the respondents to see multiple views of the watches. For the (hypothesized) lower level of craft, we used a format similar to that which is the default in most conjoint analysis software. The lower-craft profiles are primarily text-based, although we did use simple graphics. Figure 3 gives examples of the stimuli (but not the animations).

**Figure 3. Varying the Craft of the Images (from Eggers, Hauser, and Selove 2016)**
(lower craft shown first, then higher craft)



| | Watch 1 | Watch 2 | Watch 3 |
|---|---|---|---|
| Watch face: | Rectangular | Round | Rectangular |
| Case color: | Gold-colored | Gold-colored | Silver-colored |
| Band: | Brown leather band | Matching metal band | Black leather band |
| Price: | $ 349.- | $ 399.- | $ 299.- |
| Best option: | ○ | ○ | ○ |

| | Watch 1 | Watch 2 | Watch 3 |
|---|---|---|---|
| Watch face: | Rectangular | Round | Rectangular |
| Case color: | Gold-colored | Gold-colored | Silver-colored |
| Band: | Brown leather band | Matching metal band | Black leather band |
| Price: | $ 349.- | $ 399.- | $ 299.- |
| Best option: | ○ | ○ | ○ |

## Incentive-Alignment

We told the incentive-aligned respondents that 1 in 500 respondents would receive the smartwatch that he or she chose in a randomly-selected choice task as well as any difference between the displayed price and $500 (Ding 2007; Ding, Grewal, and Liechty 2005). Respondents who chose the outside option would receive $500 in cash. Incentive-aligned respondents watched a one-minute video describing the incentives (https://youtu.be/DBLPfRJo2Ho). To avoid confusing respondents, we used, for each experimental condition, a video that was consistent with that experimental condition. For example, if respondents were in the low-realism manipulation, the example choice sets shown in the video used lower-realism text-based images. In this way we measure the incremental impact of incentive alignment. Respondents who were not incentive-aligned were told that 1 in 500 would receive a cash bonus of $500. The $500 was not tied to their choices.

## Training Video

Respondents assigned to the training video were required to watch a two-minute animated video describing the smartwatch category, the smartwatch attributes, and the choice tasks (https://youtu.be/oji_bw_oxTU). We matched the videos to the experimental cells. We chose not to force an equivalent two-minute delay for respondents who were not assigned to the training video, because such a delay does not represent common practice and might introduce a demand artifact. Our goal was to determine whether or not a training video is higher craft or lower craft. It might be higher craft because well-instructed respondents might understand the task better; it might be lower craft if the additional instructions do not overcome the potential for respondent wear-out.

### Ceteris Paribus Instructions

CBC formats assume that all product attributes that are not varied in the choice tasks are held constant in the choice tasks (Green and Srinivasan 1990). If respondents do not understand that such attributes are to be held constant, they might infer unobserved characteristics to be correlated with the attributes that are varied. For example, without ceteris paribus instructions, respondents might be more likely to infer that quality changes if prices change. We used the following instructions for respondents assigned to the (hypothesized) higher level of craft. Respondents in the (hypothesized) lower level of craft received no reminders to hold all other attributes constant.

> *Please assume that all watches are from your preferred brand [adjusted to consumer's brand preference] and are compatible with your smartphone so that they can show incoming messages or calls. Assume that all of these watches have a battery that lasts a day or more, a heart rate monitor, Bluetooth, high definition color LED touchscreen, 1.2 GHz processor, 4 GB storage, and 512 MB RAM.*

## The Empirical Effect of Craft on Accuracy and Precision

The final sample of respondents who completed both waves of the study was 1,044 respondents split roughly equally among the experimental conditions in the orthogonal half replicate of the $2^4$ design. (109 respondents always chose the outside option and were eliminated as not interested in smartwatches, at least not interested in the smartwatches in the conjoint design. The elimination of respondents was not correlated with any experimental manipulation.)

Roughly 500 respondents were assigned to each condition, e.g., realistic images vs. text-based images. Respondents took approximately 200 seconds to answer the choice tasks and this did not vary among experimental manipulations. The total survey took approximately 400–500 seconds to complete. Those respondents who were assigned to incentive alignment took approximately 69 seconds longer. Those respondents who were assigned to the training video took approximately 142 seconds longer. The longer duration corresponds to the length of the mandatory videos in these conditions. In approximately 25% of the choice tasks, respondents chose the no-choice outside option. This was slightly higher (3–5%) for more-realistic images, for those who saw the training video, and for those who were incentive aligned.

### Impact of Craft on Accuracy

Neither incentive alignment, the training video, nor ceteris paribus instructions affected the relative part-worths substantially. For example, on average, differences in relative attribute importances between manipulations varied by but a few percentage points, usually by no more than one percentage point. On the other hand, realistic images mattered. Those respondents who were shown more realistic images were more likely to value differences in the watch band (40% vs. 27% relative importance of watch band) and less likely to value differences in color (17% vs.

22% relative importance of watch-face color). They were also relatively less price sensitive (21% relative importance of price vs. 28% relative importance of price). Thus, it appears that only the realism of the images affects accuracy in our data.

Accuracy (the relative part-worths) affects WTP, where WTP is interpreted as consumer surplus—the demand curve. Although, in McFadden's (2014) formulation, the relative part-worths are the WTPs, there are issues with this interpretation when hierarchical Bayes or empirical Bayes analysis is used. First, the part-worths are heterogeneous. This heterogeneity is inherent whether we sample from the hyper-distribution, from the full distribution of respondent-level part-worths, or if we use the mean part-worths for each respondent. Second, the conjoint study only collects data within certain price ranges. For example, our prices varied from $299 to $449. Samples from the hyper-distribution, samples within the full distribution, or even the mean part-worths might imply WTPs outside this range. It would violate sound scientific principles to extrapolate beyond the price range that was used in the survey. Thus, we need a robust summary that does not violate these scientific principles. One method is to use robust statistics, such as medians. Another procedure that is commonly applied is to use a two-product simulator in which one product has the attribute level of interest and the other has the lower level of the attribute. The price that equalizes the predicted shares of both products can then be interpreted as WTP. We adopt this common practice for estimating the market's WTP.

Because only the realism of the images affects accuracy, only the realism of the images affects WTP substantially. As shown in Table 3, this impact can be dramatic suggesting that researchers should pay close attention to the realism of the images. The default of text-based formats may underestimate the willingness to pay for a product attribute.

**Table 3. The Realism of Images in Conjoint Analysis Affects Estimated WTP**

| Calculated vis the Simulation Method | More Realistic Images | Text-Based Images |
|---|---|---|
| **Round to Rectangular Face** | $103 | $39 |
| **Gold to silver color** | $65 | $59 |
| **Brown to black band** | $130 | $42 |
| **Metal to black band** | $132 | $4 |

## Impact of Craft on Estimation Precision ($\gamma^{estimation}$)

The realism of the images and incentive alignment do not impact the precision that we estimate from the estimation data, $\gamma^{estimation}$. This is intuitive, $\gamma^{estimation}$ summarizes the internal consistency of the estimated logit model. When respondents are asked to choose among text-based descriptions, they might be extremely consistent in reporting their preferences for text-based descriptions. This does not necessarily mean that the reported preferences for text-based stimuli describe how respondents would react in the marketplace.

Two aspects of craft affect estimation precision negatively. Estimation precision is significantly lower for respondents who saw the training video and those who were provided with the ceteris paribus instructions. The training video effect is intuitive. The extra time necessary to watch the training video may have been burdensome. It appears the time to watch the video was sufficiently burdensome so that respondents were less internally consistent in their choices in the estimation choice sets. The ceteris paribus instructions had a small, but statistically significant, negative impact on precision.

### Impact of Craft on External Validity Precision ($\gamma^{externally\ adjusted}$)

When adjusted for external validity, precision was significantly better for more realistic images (vs. text-based images) and for incentive alignment (vs. no incentive alignment). For our data, these results imply that higher craft means using enhanced image realism and incentive alignment. Providing a training video or ceteris paribus instructions did not affect the adjustment based on external validation. However, overall the externally-adjusted precision in the training video condition is lower due to the lower estimation precision. Although significant, the effect is rather low in magnitude. Overall, there is no effect of ceteris paribus instructions; either they have little effect in general or respondents instinctively held all other attributes constant when making choices among the product profiles in our research setting.

The training video lowered the externally adjusted precision. For our data, it appears that the negative wear-out effect was larger than the (hypothesized positive) training effect. While our training video was professional quality, it was long relative to the total time of the questionnaire (142 incremental seconds out of approximately 500 seconds). We hypothesize that more concise training videos might have a more positive impact. Training videos might also be valuable for product categories that are harder for consumers to grasp. At minimum, our results caution conjoint analysis analysts that training videos must be crafted and pretested carefully.

## Craft Affects Managerial Recommendations because Craft Affects Precision

We have already seen that craft in the realism of images affects accuracy and, by implication, estimated WTP. Craft also affects precision. We focus on craft in incentive alignment because craft in incentive alignment has a substantial effect on precision but not accuracy. By focusing on incentive alignment, we can illustrate a precision effect that is not confounded with an accuracy effect. (We return to realism in the images, and other aspects of craft, later in this paper. Image realism illustrates nicely the joint effect of precision and accuracy. No interactions among different aspects of craft were observed.)

### Impact on Predicted Equilibrium Prices and Profits

Table 4 compares estimated equilibrium prices and profits (shares*price). The results are based on part-worths that are externally adjusted. The market is a two-product market in which the products vary on only the color of the watch face. The estimated equilibrium prices and profits are for the products with the corresponding color.

In Table 4, improved craft predicts lower (and presumably more-correct) prices and profits. Lower craft (no incentive alignment) appears to overstate prices by about 3% and profits by about 5%. We obtain comparable effects for the shape of the smartwatch face and the type of smartwatch band. Externally adjusting the part-worths tends to harmonize the results. These differences due to craft are substantially larger when relying on estimation precision only, which is addressed next.

**Table 4. The Effect of Craft (Precision) on Equilibrium Prices and Profits**

| | Equilibrium Prices | | Equilibrium Profits | |
|---|---|---|---|---|
| | Incentive Alignment | No Incentive Alignment | Incentive Alignment | No Incentive Alignment |
| **Gold-colored Smartwatch** | $311 | $328 | $99 | $106 |
| **Silver-colored Smartwatch** | $343 | $347 | $124 | $129 |

## External vs. Internal Precision Matters When Calculating Predicted Equilibrium Prices and Profits

Recall that craft in the realism of images had a two-fold effect; lower craft lowered both accuracy and precision. Table 3 illustrated that less-realistic images lowered predicted WTP for all smartwatch attribute-level differences—mostly because lower craft increased the estimated relative importance of price. For example, lower craft lowered the estimated WTP for gold vs. silver-colored smartwatches from $65 to $59—about 9%. In contrast, a lower (externally-adjusted) precision suggests lower price sensitivity and higher equilibrium prices (e.g., see Table 2). The effects of accuracy (WTP) and precision (price sensitivity) might counteract one another.

We put together the joint effect of accuracy and precision and compare the differences in predicted equilibrium prices. For estimation precision, we expect the effect on accuracy to dominate. This is shown in Table 5a. Predicted equilibrium prices are roughly 25% lower for text-based images versus realistic images. Predicted equilibrium profits are 14% lower.

For externally-adjusted precision, text-based images lowered precision significantly, an effect that might counteract the effect of the lower accuracy (i.e., higher price sensitivity). This joint effect of accuracy and precision is shown in Table 5b. Predicted equilibrium prices and profits are, on average, higher for text-based images, but only by 6% and 11% respectively. (Note that predicted equilibrium prices and predicted equilibrium profits are higher in both conditions when based on externally adjusted precision vs. estimation precision. They increase by 47% and 26%, respectively.)

In our data, lower accuracy and lower precision counteract one another for craft based on the realism of the images. But that will not always be the case. If lower craft were to lower the relative importance of price, then the two effects would reinforce one another. Lower craft would have an even bigger effect on managerial recommendations and on patent or copyright valuations. The key message in Table 5 is that researchers should pay attention to craft *and* to external-validity adjustments to precision.

**Table 5. Adjustments to Reflect External Validity Matter Managerially**
**(Example Where Decreased Accuracy and Decreased Precision Counteract)**

| Table 5a | Equilibrium Prices | | Equilibrium Profits | |
|---|---|---|---|---|
| **Results for Estimation Precision** | More Realistic Images | Text-Based Images | More Realistic Images | Text-Based Images |
| **Round Smartwatch Face** | $233 | $194 | $79 | $78 |
| **Rectangular Smartwatch Face** | $317 | $219 | $124 | $96 |

| Table 5b | Equilibrium Prices | | Equilibrium Profits | |
|---|---|---|---|---|
| **Results for Externally-Adjusted Precision** | More Realistic Images | Text-Based Images | More Realistic Images | Text-Based Images |
| **Round Smartwatch Face** | $298 | $350 | $92 | $117 |
| **Rectangular Smartwatch Face** | $386 | $377 | $134 | $134 |

### Impact on the Strategic Recommendations for the Attribute Levels of a Product

The description of the equilibrium in attributes and prices is beyond the scope of this paper. For more details on the theory, see Hauser, Eggers, and Selove (2016). They show for the magnitudes of precision that we find in our empirical data, the innovator would be advised to launch the more preferred silver-colored smartwatch. The follower's actions depend upon the precision that the follower estimates for the market. In particular, a follower would be advised to launch a:

- Gold-colored smartwatch if the true precision were higher
- Silver-colored smartwatch if the true precision were lower

We get similar effects for the watch face and the watch band.

## INCREASED SAMPLE SIZE DOES NOT COMPENSATE FOR LOWER CRAFT

It is tempting to equate precision, as defined in this paper, with precision of the part-worth estimates. This is an incorrect interpretation. Precision (scale of the part-worths) is a different concept than the standard errors of the estimates of the part-worths, or the related concepts in Bayesian analysis.

To illustrate that they are different concepts, we re-estimated all of our models using a randomly selected 50% sample. On average, the standard errors of the estimates of $\gamma^{estimation}$ and $\gamma^{externally\ adjusted}$were 32% lower for the full sample compared to the random half sample, but the magnitudes of the estimates were comparable. Averaged over all conditions, the estimation precisions were within less than 1%, and the externally-adjusted precisions within 3%, when we compare estimates based on the full sample to estimates based on a random half of the sample.

Sample size does not overcome lower craft!

## Discussion and Summary

### Theory

Accuracy matters, but its effect has been well-studied. Accurate relative part-worths are important for deciding which attribute levels to include in products and for calculating willingness-to-pay (WTP) as input to other managerial recommendations (and in litigation, as input to damages experts). However, precision (scale) matters as well. Precision directly affects predictions of equilibrium prices and profits and, as a result, recommendations on the *strategic* selection of attribute levels for products. The market reacts according to true precision ($\gamma^{true}$), but analysts make recommendations based on estimated precision. Recommendations vary dramatically depending upon whether the recommendations are based on estimation precision ($\gamma^{estimation}$) or externally-adjusted precision ($\gamma^{externally\ adjusted}$).

### Craft

Craft affects both accuracy and precision. For the situation we examined empirically, we found:

- More realistic images increased both accuracy and precision.
- Incentive alignment increased precision, but had little effect on accuracy.
- Training videos had no effect on accuracy, but appear to decrease both estimation precision and externally-adjusted precision.
    - In our data, the training videos were time-consuming for consumers to watch and may have led to respondent wear-out. The wear-out effect might have been stronger than the training effect.
    - Well-designed training videos might enhance precision. One must craft and pretest such videos carefully.
- Ceteris paribus instructions had no substantial effect on either accuracy or precision. This result might be due to the fact that we used instructions that mimicked the status quo on the market, which consumers might have assumed implicitly even without instructions.
- Increased sample size does not compensate for lower craft.

### Managerial Recommendations

Craft affects managerial recommendations. For the situations we examined empirically, we found:

- The realism of the images changed the relative importance of the attributes and decreased the relative importance of price. WTPs were higher for more realistic images.
- Predicted equilibrium prices and profits depend upon craft and whether or not the part-worths are adjusted for external validity.

- Managerial recommendations may change if they are made using precisions as estimated from the conjoint analysis data or if they are made using precisions adjusted for external validity.
    - In some cases, the effect of craft on accuracy and precision counteract one another. In other cases, the effects of craft on accuracy and precision reinforce one another.
    - Because the directional impact cannot be predicted ahead of time, higher craft is advised.
- The correct strategic selection of attribute levels for products depends upon accuracy and precision, and, hence, craft.

## Summary

Craft matters! High craft avoids making the wrong (or costly) managerial recommendations. High craft avoids estimating incorrect valuations for copyrights and patents.

Felix Eggers

John R. Hauser

Matthew Selove

## References

Allenby, Greg M., Jeff Brazell, John R. Howell, and Peter E. Rossi (2014), "Valuation of Patented Product Features." *Journal of Law and Economics*, 57:3 (August). 629–663.

Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28 (September), 273–83.

Brazell, Jeff D., Christopher G. Diener, Ekaterina Karniouchina, William L. Moore, Válerie Séverin, and Pierre-Francois Uldry (2006), "The No-Choice Option and Dual Response Choice Designs." *Marketing Letters*, 17(4), 255–268.

Cameron, Lisa, Michael Cragg, and Daniel McFadden (2013), "The Role of Conjoint Surveys in Reasonable Royalty Cases," *Law360*, October 16.

Ding, Min (2007), "An Incentive-aligned Mechanism for Conjoint Analysis." *Journal of Marketing Research*, 54, (May), 214–223.

Ding, Min, Rajdeep Grewal, and John Liechty (2005), "Incentive-aligned Conjoint Analysis." *Journal of Marketing Research* 42, (February), 67–82.

Eggers, Felix, John R. Hauser, and Matthew Selove (2016), "Precision Matters: How Craft in Conjoint Analysis Affects Price and Positioning Strategies," (Cambridge, MA: MIT Sloan School of Management).

Green, Paul E., and V. Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice." *Journal of Marketing*, 54, 4, (October), 3–19.

Hauser, John R., Felix Eggers, and Matthew Selove (2016), "The Strategic Implications of Precision in Conjoint Analysis," (Cambridge, MA: MIT Sloan School of Management).

McArdle, Meghan (2000), "Internet-based Rapid Customer Feedback for Design Feature Tradeoff Analysis," S. M. Thesis. (Cambridge MA: MIT Sloan School of Management).

McFadden, Daniel L. (2014), In the Matter of Determination of Rates and Terms for Digital Performance in Sound Recordings and Ephemeral Recordings (WEB IV) Before the Copyright Royalty Board Library of Congress, Washington DC, Docket No. 14-CRB-0001-WR, October 6.

Meissner, Martin, Harmen Oppewal, and Joel Huber (2016), "How Many Options? Behavioral Responses to Two versus Five Alternatives per Choice," *Proceedings of the 19th Sawtooth Software Conference,* Park City, UT, September 26–20.

Mintz, Howard (2012), "2012: Apple Wins $1 Billion Victory over Samsung," *San Jose Mercury News*, August 24.

Orme, Bryan K. (2010), *Getting Started with Conjoint Analysis: Strategies for Product Design and Pricing Research, 2E*. (Research Publishers LLC: Madison WI).

Sawtooth Software (2009). http://www.sawtoothsoftware.com/support/technical-papers/hierarchical-bayes-estimation/cbc-hb-technical-paper-2009.

Schaeffer, Nora Cate and Stanley Presser (2003), "The Science of Asking Questions," *Annual Review of Sociology*, 29, 65–88.

Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22, 3, (Summer), 273–303.

Wlömert, Nils and Felix Eggers (2016), "Predicting New Service Adoption with Conjoint Analysis: External Validity of BDM-Based Incentive-Aligned and Dual-Response Choice Designs," *Marketing Letters*, *27*(1), 195–210.

# How Many Options? Behavioral Responses to Two versus Five Alternatives per Choice

**Martin Meissner**
*University of Southern Denmark/Monash University*
**Harmen Oppewal**
*Monash University*
**Joel Huber**
*Duke University*

## Abstract

What is an appropriate number of alternatives per choice task? Why are two or five alternatives so rarely used? We characterize the contexts where Choice-Based Conjoint (CBC) on pairs makes sense when projecting to real decisions and use eye-tracking to study how respondents search for information when answering choice tasks with either two or five alternatives.

## Introduction

While there has been substantial work asking how many choice tasks are needed in a CBC study, less attention has been paid to determining the appropriate number of alternatives per choice task. The typical answer to the question of how many alternatives to include revolves around the tradeoff between greater statistical efficiency and increased task difficulty for the respondent. From the perspective of statistical efficiency (as tested using computer simulations) paired comparison choice tasks produce less efficient designs (Bunch, Louviere and Anderson 1996; Louviere and Woodworth 1983). Increasing the number of alternatives in each task provides greater statistical efficiency, because multinomial logit models in fact assume that each final choice is based on a comparison of the selected alternative to all of the available options. However, answering more complex choice questions also makes the choice task more difficult for the respondents and makes it more likely that respondents use simplifying decision heuristics (Bettman, Johnson and Payne 1991; Todd 2007).

As a starting point for our research, we investigated how frequently practitioners who published in the past four Sawtooth Software Proceedings used two, three, four or five alternatives in their CBC studies. We found about 20 studies in each of the four Proceedings. We were surprised by a trend in the last few years which suggests substantial changes in practice. Examining Proceedings in 2010, 65% of the studies mentioned had 5 or more alternatives per choice while 35% had 3 or 4. By 2015 that proportion had reversed, with 10% choosing 5 or more, and almost 90% choosing 3 or 4. Across time, the proportion of studies using two alternatives has consistently stayed under 10%.

The paper by Pinnell and Englert (1997) may be one reason why pairs rarely appear in the recent Sawtooth Software Proceedings. The authors varied the number of alternatives in three experiments and concluded that respondents are capable of accurately answering choice tasks with up to seven options. Compared with two alternatives the authors found that it took about 33% more time for respondents to evaluate four alternatives and about 60% more time to evaluate seven alternatives. The authors concluded that it is advisable to use more than two options in a choice task because in their studies pairs had lower predictive validity, were less stable and did not save much time relative to choice tasks with more alternatives.

We propose that the performance of pairs should be reevaluated in light of respondents' behavioral responses and that responses to choice tasks depend on how respondents search for information when making their choices. We use eye-tracking to investigate how respondents allocate their attention in CBC choice tasks with two (pairs) or five alternatives (quints). We focus on pairs and quints, because from a practical perspective they reflect the range in the number of alternatives reported in recent Sawtooth Software Proceedings.

We investigate task differences from three different perspectives: First, we examine the ways respondents search for decision-relevant information in pairs and quints, focusing on processing patterns and ways that respondents learn to more efficiently complete their task. Second, we assess to what extent respondents perceive the task as difficult. Third, we compare pairs versus quints in terms of internal consistency and their ability to predict holdout choices from triples.

## Behavioral Responses to Choice Complexity

Results from previous studies which have investigated information processing suggest that in choice tasks including only two alternatives respondents use compensatory decision strategies, such as the additive difference strategy (ADD, see e.g., Payne 1976). In line with this research, we expect that in pairs respondents process almost all attribute information that is available (Payne *et al.* 1992) and compare the two alternatives in a step-by-step, attribute-wise and top-to-bottom manner (Russo and Dosher 1983). Using a systematic and complete search process, respondents focus fairly evenly across all attributes, including the less important ones.

In contrast, in more complex choice tasks, respondents have been shown to simplify their choice (Bettman, Johnson and Payne 1991; Ford *et al.* 1989; Payne *et al.* 1992). Compared to a task consisting of pairs, in a task with larger numbers of decision alternatives, respondents can be expected to display greater cognitive effort, but at the same time also show a greater degree of simplification. Respondents can, for example, simplify their search by eliminating alternatives from further consideration based on important attributes. Process tracing studies have shown that respondents often use heuristics, such as the lexicographic rule, to reduce the number of options by excluding those not meeting a minimum level for a particular attribute (Payne *et al.* 1992). We also expect that respondents simplify their search in later tasks because they will learn from the earlier choice tasks which attributes matter most to them (Meissner and Decker 2010; Orquin, Bagger and Mueller Loose 2013).

## Study Design

Upon arrival in the laboratory, participants received a general instruction and were familiarized with the eye-tracker. The main task consisted of a self-guided computer-based conjoint survey. It first introduced vacation packages as the product of interest. Next it asked respondents about their prior purchase experience, future purchase intention as well as purchase familiarity and involvement regarding vacation packages. The following screens then explained the attributes and levels of the vacation packages. Respondents then answered eight choice tasks that were presented on separate screens. The profiles in each task were randomly generated with Sawtooth Software's complete enumeration algorithm (Orme 2013). The statistical strength of the designs for pairs and for quints was determined based on eight choice tasks, 40 respondents, and using Sawtooth Software's complete enumeration algorithm. The relative D-efficiency ratio of the pairs vs. quints design is 174.4/133.6=1.31. This result means that statistically, 31% more observations are needed for pairs to achieve the same aggregate logit efficiency as quints.

After the sequence of eight choice tasks, respondents were asked about their search goals and perceived task difficulty, their frustration and the perceived similarity of options for the last of their eight choice tasks. Finally, respondents answered two holdout choice tasks consisting of three randomly generated alternatives (triples). The survey ended with socio-demographic questions. The results of these additional survey questions are beyond the scope of the current paper and are not discussed.

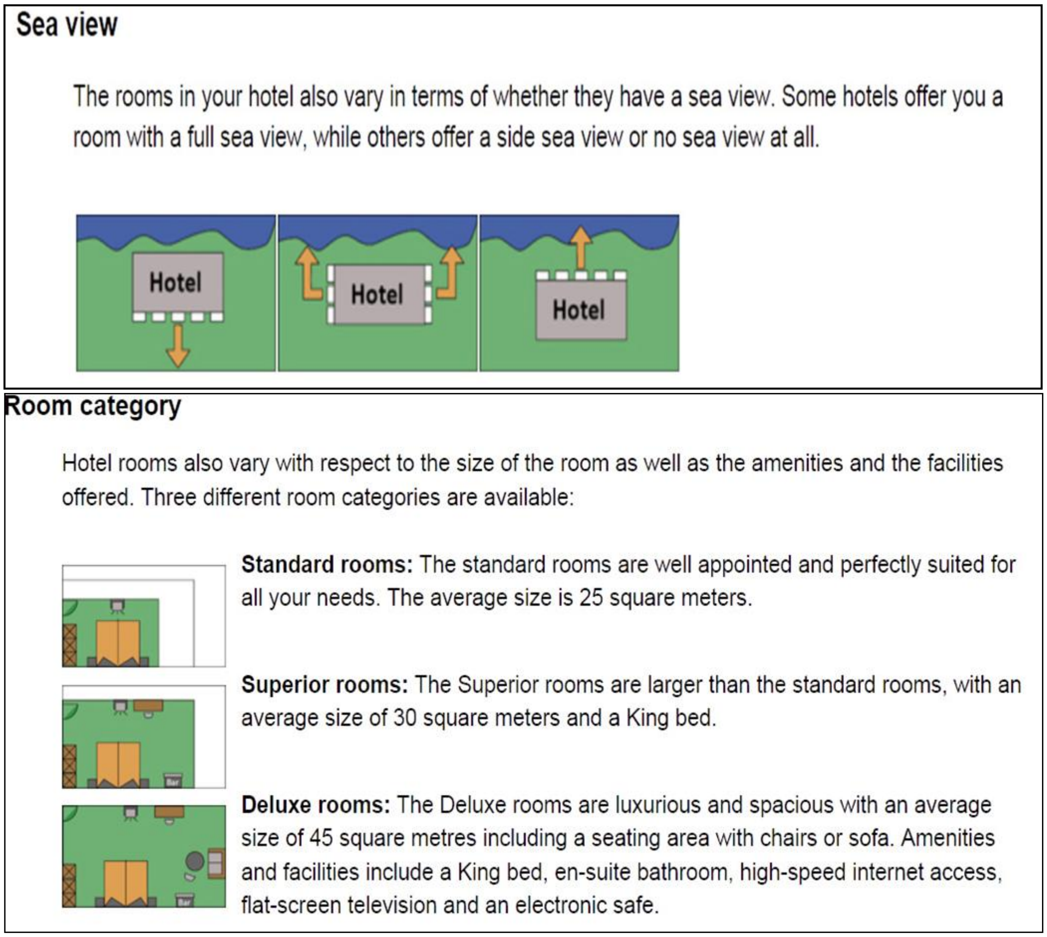
The hypothetical vacation packages were characterized by six attributes, each with three levels. As common in CBC studies, the attributes appeared in a fixed display order. Table 1 describes the attributes and their levels.

**Table 1. Vacation Package Attributes and Levels**

| Attributes | Attribute levels | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| Food quality | good | very good | excellent |
| Customers recommending | 50% | 70% | 90% |
| Distance to CBD | 3km | 2km | 1km |
| Sea view | no sea view | side sea view | full sea view |
| Price per person | $899 | $799 | $699 |
| Room category | standard | superior | deluxe |

Separate screens helped respondents become familiar with each of the attributes and their levels before respondents answered the choice tasks. Figure 1 gives an example for the attributes "sea view" and "room category." We agree with Eggers, Hauser and Selove (2016, this Proceedings) that craft is important in designing a preference measurement study and used images and ceteris paribus instructions to enhance precision and accuracy. We did not use training videos and did not incentive-align our respondents but do not believe our results comparing pairs vs. quints would change if we included those changes.

**Figure 1. Example Instructions Explaining the Attribute Levels of the Attributes "Sea View" (Top) and "Room Category" (Bottom)**



Sea view

The rooms in your hotel also vary in terms of whether they have a sea view. Some hotels offer you a room with a full sea view, while others offer a side sea view or no sea view at all.

Room category

Hotel rooms also vary with respect to the size of the room as well as the amenities and the facilities offered. Three different room categories are available:

**Standard rooms:** The standard rooms are well appointed and perfectly suited for all your needs. The average size is 25 square meters.

**Superior rooms:** The Superior rooms are larger than the standard rooms, with an average size of 30 square meters and a King bed.

**Deluxe rooms:** The Deluxe rooms are luxurious and spacious with an average size of 45 square metres including a seating area with chairs or sofa. Amenities and facilities include a King bed, en-suite bathroom, high-speed internet access, flat-screen television and an electronic safe.

The eye-tracking study was carried out at Monash University (Australia). A total of 39 respondents finished the questionnaire with pairs and 38 respondents finished the questionnaire with quints.

## Information Processing and Eye-Tracking

Eye-tracking is one of the most reliable approaches for observing humans' attentional processes.[1] Modern eye-tracking systems use video images of the eyes to determine the so-called "point of regard." It reveals that human perception is primarily based on two states of the eyes: fixations and saccades. A fixation is defined as a state where the eyes are relatively stable and "rest on" a certain stimulus. A rapid movement of the eyes between two consecutive fixations is called a saccade. Typically a fixation is between 100 and 500 milliseconds (ms) long with an average of about 250 ms. The fixation duration largely depends on the viewed stimuli and their characteristics (Rayner 1998). A saccade typically lasts between 30 and 50 ms. Studies have shown that humans cannot acquire information during a saccade (Rayner 1998) because the brain

[1] The interested reader is referred to Holmqvist *et al.* (2011) for a more comprehensive introduction to eye-tracking and for a discussion of adequate measures.

blocks visual processing during eye movements in a way that neither the motion of the eye nor the gap in visual perception is noticeable to the individual.

A Tobii T120 recorded the eye movements in our study. This system has an accuracy of 0.4° of visual angle and a sampling rate of 120 Hz. The infrared sensors are built into a 17" thin-film transistor (TFT) monitor with a resolution of 1280 x 1024 pixels. A standard 9-point calibration routine was used to calibrate participants' eye movements (Tobii Software 2016). When placing the respondents in front of the eye-tracker, we made sure that the distance indicator provided by the Tobii software displayed a value between 50 and 80 cm (ideally 60 cm) as recommended by the Tobii handbook. Respondent answers were given solely with a computer mouse.

The areas of interest were defined as cells in the display matrix; they were all of equal size, non-overlapping, and the number ranged between 12 (2 alternatives * 6 attributes) and 30 (5 alternatives * 6 attributes) cells. Fixations were defined as continuous gazes within each area of interest. We used the standard Tobii fixation filter to determine fixations (Tobii Software 2016).

Moreover, it is important to emphasize that we used only simple text labels to describe the features (see Figures 2 and 3). It is therefore unlikely that the feature stimuli differed regarding their saliency, which could have produced differences with respect to the number of fixations to features (effects of bottom-up attention). All respondents reported to have normal or corrected to normal vision. In order to simplify the analyses, we only used the data of the right eye. However, the results do not differ if we use the data for the left eye or the average of the left and the right eye as provided by the Tobii software.

## Results

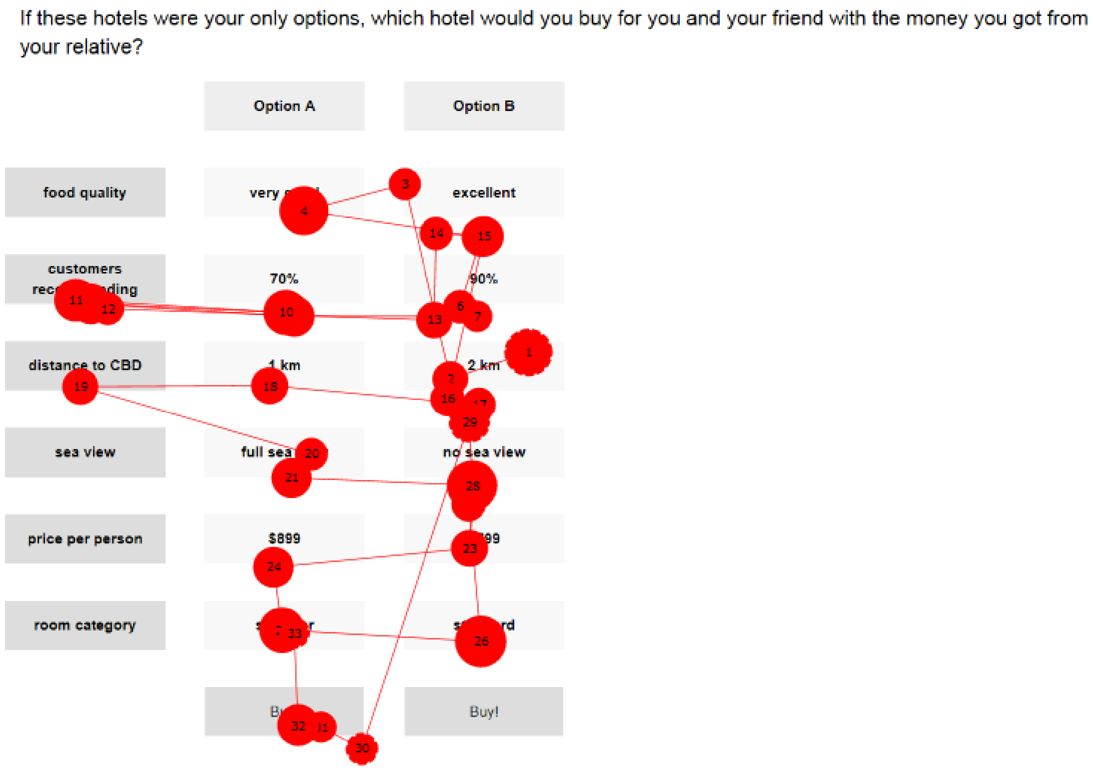
### Observing What Respondents Do

Before analyzing respondents' search patterns using known measures of information search, we visually inspected the scanpaths of the choice tasks for every respondent. The inspection of the videos showed that the search patterns often matched our expectations as outlined above in the section "Behavioral Responses to Choice Complexity." Two example paths of fixations that are easily interpreted are depicted in Figure 2 for a pairs task and in Figure 3 for a quints task. We encourage the reader to watch the corresponding videos which are available on YouTube in fast (https://youtu.be/wmpy7O-dZFY) and slow (https://youtu.be/9YZBVyI9TZM) motion for pairs and in fast (https://youtu.be/qXZYIz8eEdc) and in slow (https://youtu.be/he9SjPYVP8Q) motion for quints.

The search process in Figure 2 for pairs follows a typical additive difference model. After two initial fixations to the center of the screen, the respondent starts the search by looking at the top attribute "food quality." The search continues by comparing the two options with respect to each attribute, moving from the top to the bottom of the screen. In this example all attribute levels are fixated at least once. After having fixated the last attribute, the respondent checks three attribute levels of option B before then choosing option A. The respondent possibly is reassured by the undesired aspects of option B before making the final decision. It is also interesting to see that this respondent does not look at the question text, in this case because she has seen other choice tasks before. She also looks at the description of the attributes only two times, i.e., when comparing the alternatives with respect to "customers recommending" and "distance to CBD."

We speculate that she is looking at the attribute labels because the attribute levels are not self-explanatory, as it is not clear what “70%” means without looking at the attribute label.

**Figure 2. An Example Path of Fixations for Pairs**

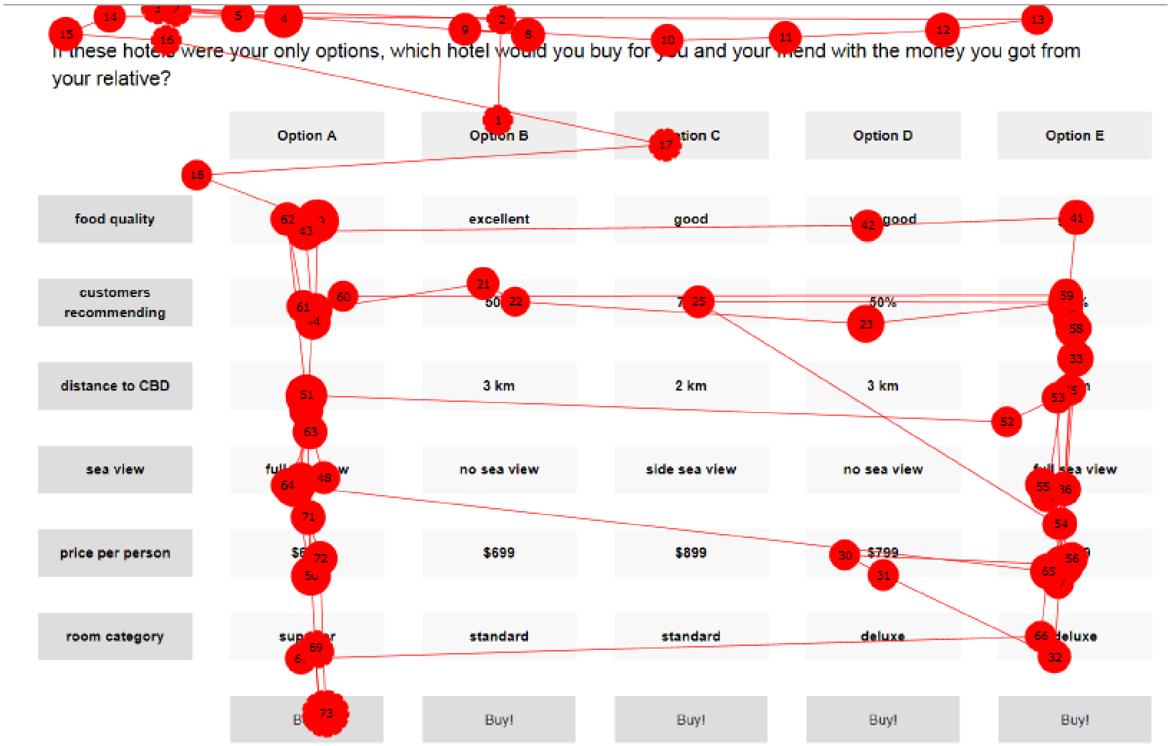


If these hotels were your only options, which hotel would you buy for you and your friend with the money you got from your relative?

| | Option A | Option B |
|---|---|---|
| food quality | very good | excellent |
| customers recommending | 70% | 90% |
| distance to CBD | 1 km | 2 km |
| sea view | full sea view | no sea view |
| price per person | $899 | [illegible]99 |
| room category | superior | standard |
| | Buy! | Buy! |

The example search process for quints shown in Figure 3 is quite different. In this case, the respondent starts the task by reading the question text. Next, the respondent looks at the attribute “customers recommending” and compares all five alternatives with respect to that attribute. Only options A and E have a customer recommending rating of 90% in this choice task. The customer rating seems to be most important for the respondent and that is why she probably starts the search process by looking at that particular attribute. After identifying options A and E as promising the respondent’s search process changes substantially. The respondent evaluates option E in detail by looking at all the attribute levels of that alternative. The respondent then jumps to option A, which is also evaluated in detail. In what follows, we can see that the respondent focuses only on these two options, A and E, by going back and forth between them. Options B, C and D are only fixated incidentally, perhaps because they are in the way. Many of the levels for the attributes “food quality,” “distance to CBD,” “sea view” and “room category” are not fixated at all for these three options. In all, the search process can be best described as a staged simplification process. In the first step the respondent uses one attribute to identify promising alternatives, in a second step the remaining alternatives are evaluated holistically and compared to one another. We can also see that the search process in the second step includes more transitions within alternatives compared to the search process for pairs.

## Figure 3. An Example Path of Fixations for Quints



If these hotels were your only options, which hotel would you buy for you and your friend with the money you got from your relative?

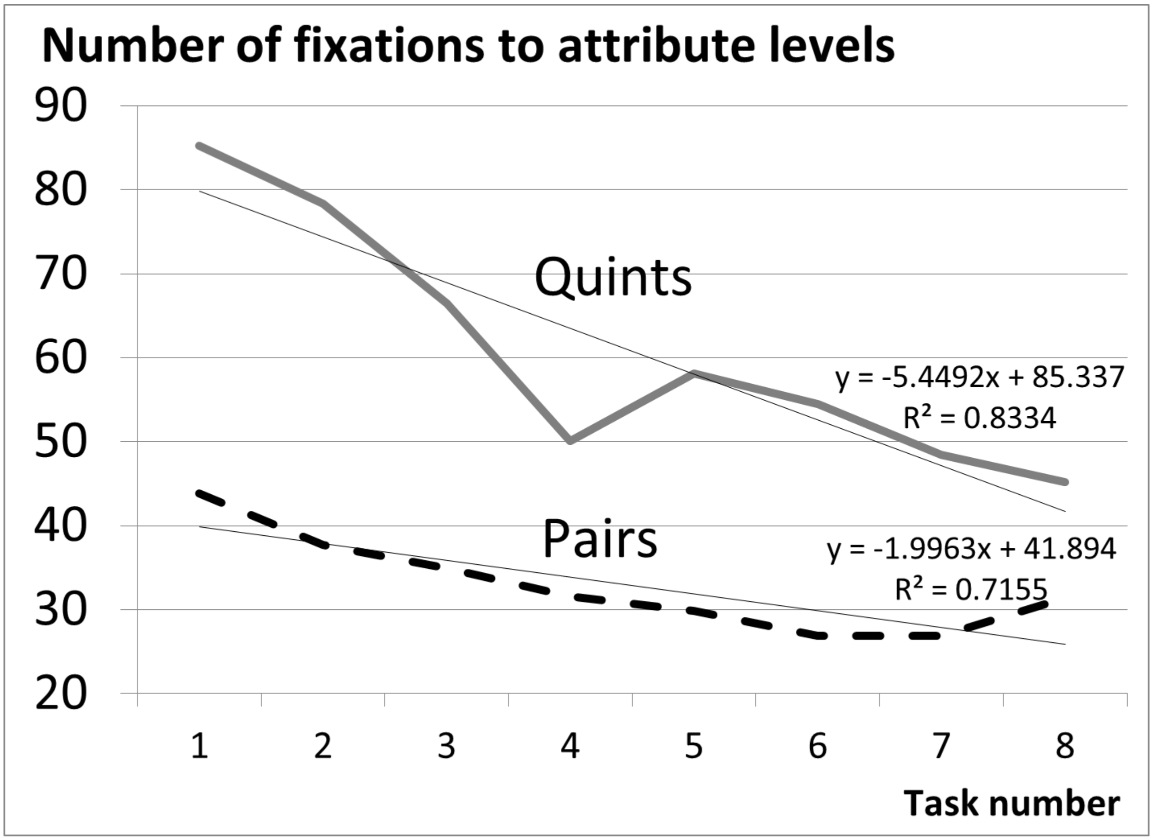
| | Option A | Option B | Option C | Option D | Option E |
|---|---|---|---|---|---|
| food quality | [illegible] | excellent | good | very good | [illegible] |
| customers recommending | [illegible] | 60% | [illegible] | 50% | [illegible] |
| distance to CBD | [illegible] | 3 km | 2 km | 3 km | [illegible] |
| sea view | full sea view | no sea view | side sea view | no sea view | full sea view |
| price per person | [illegible] | $699 | $899 | $799 | [illegible] |
| room category | superior | standard | standard | deluxe | deluxe |
| | Buy! | Buy! | Buy! | Buy! | Buy! |

These examples are chosen because they cleanly illustrate the use of expected decision rules. In fact we found that the processes followed many patterns that changed within as well as across respondents. Next we show how quints and pairs differed in using five general measures of information processing: the number of fixations, percent of information accessed, frequency of within vs. between attribute transitions, top-down vs. bottom-up order processing, and average duration of each fixation.

## Information Processing

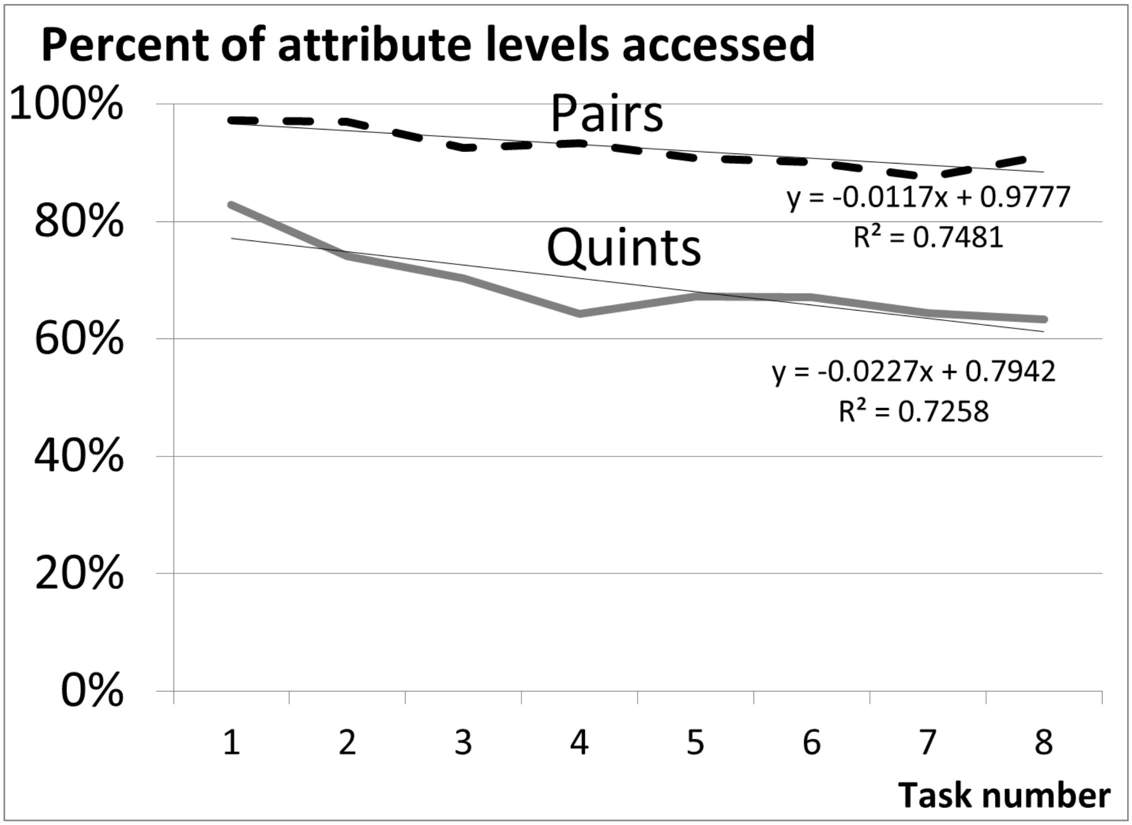
First, we investigated the number of fixations.  Changing the number of alternatives had a substantial impact on the number of fixations required for each choice. As shown in Figure 4, respondents in choice tasks comprising five alternatives expend about twice as many fixations (M=60.82, SD=39.85) to make a decision than those encountering sets with two alternatives (M=32.91, SD=18.40). This difference is highly significant. Further, as respondents become more experienced with the task they expend fewer fixations. Respondents in the quints condition adapt faster. The number of fixations dropped about 27% from the first to the last choice task for pairs. For the quints the drop is about 43%. This result replicates Pinnell and Englert’s (1997) finding that respondents accelerate processing more in choice tasks with seven than in choice tasks with two alternatives. The result suggests that respondents largely change the way they process the information in quints by simplifying more in later choice tasks. For pairs, there is minimal simplification in later choice tasks. The observed reduction in the number of fixations is also in line with findings by Meissner, Musalem and Huber (2016) who used eye-tracking to show that respondents become more efficient; that is, they need fewer fixations and become more consistent as they progress in a decision sequence of multi-attribute choice tasks. Stüttgen, Boatwright and Monroe (2012) found a similar decrease in the number of fixations when testing choice from simulated product shelves.

**Figure 4. Number of Fixations on Attribute Levels**



Second, in order to assess the degree of simplification we investigate how many attribute levels the respondents fixated on at least once. In line with our expectations, respondents in the pairs condition accessed 92% of the information available, compared with 69% for quints. As can be seen from Figure 5, pairs access a greater proportion of information and are less likely to reduce that coverage with task experience. Our finding is in line with Yang, Toubia and De Jong (2015) who investigated a sequence of 20 choice tasks including four alternatives. Yang *et al.* found that respondents across all tasks looked at about 70% of the available information and found a similar downward trend with respect to the percent of attribute levels fixated in later tasks.

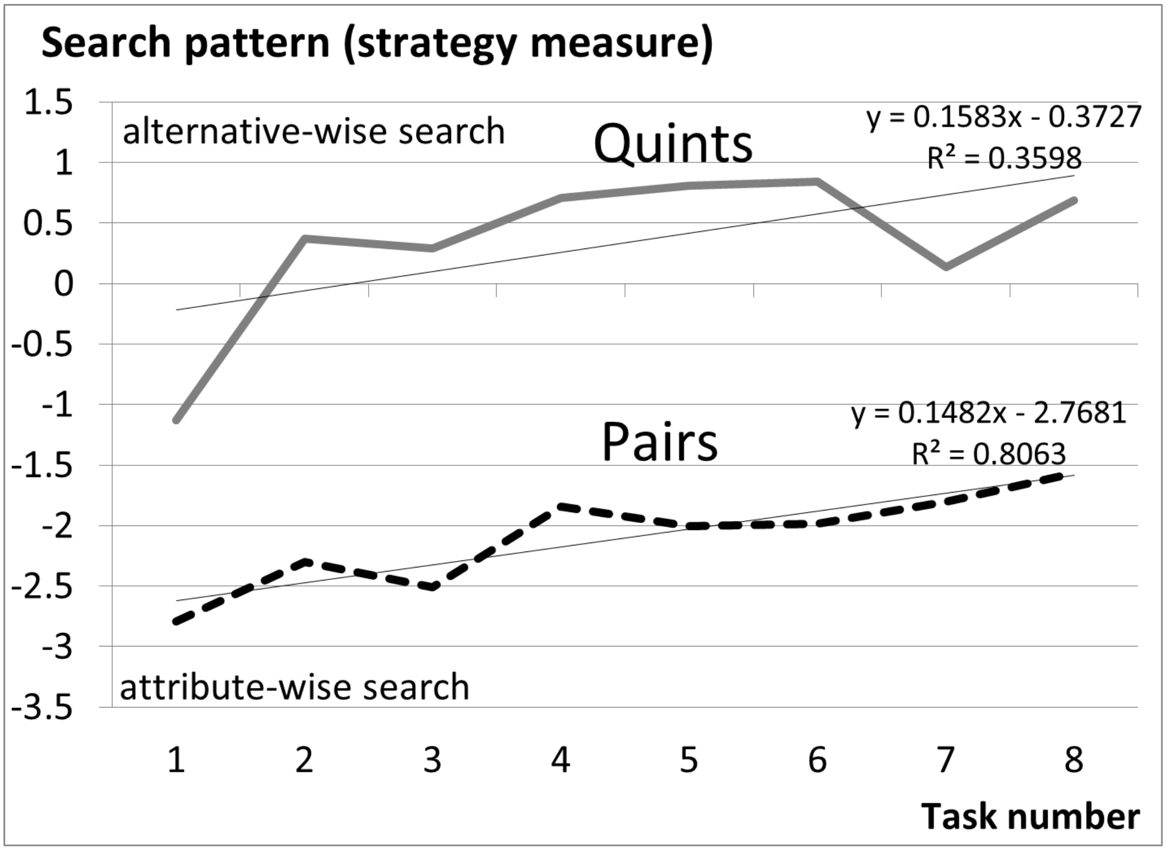
**Figure 5. Percent of Attribute Levels Accessed by Task Number**



Third, we compare the search (or saccade) pattern used in pairs and quints. A frequently used measure to describe the search pattern is the strategy measure (Böckenholt and Hynan 1994) which quantifies the extent to which information is searched attribute-wise, i.e., comparing alternatives within attributes, or alternative-wise, comparing attributes within alternatives. Because the strategy measure takes into account that the probability of attribute-wise and alternative-wise transitions changes for different numbers of alternatives, the strategy measure is the preferred index for assessing how respondents process task-relevant information (Schulte-Mecklenbeck, Kühberger and Ranyard 2011). A negative value of the strategy measure indicates attribute-wise processing whereas positive values indicate alternative-wise processing.

With respect to the search patterns we find that respondents conducted more within-attribute processing for pairs. This result therefore is in line with previous work (Russo and Dosher 1983) showing that decision makers process the information primarily attribute-wise on pairs. For quints respondents used a mixture of attribute-wise and alternative-wise processing, but in both conditions greater task experience resulted in greater alternative-wise processing. That result for both quints and pairs is consistent with Meissner and Decker (2010) who observe a progression to within-alternative transitions. In our pairs condition, however, the shift towards alternative-wise processing is small. Across all tasks respondents process the information attribute-wise which suggests that respondents continued to emphasize an additive difference strategy throughout their eight choices.

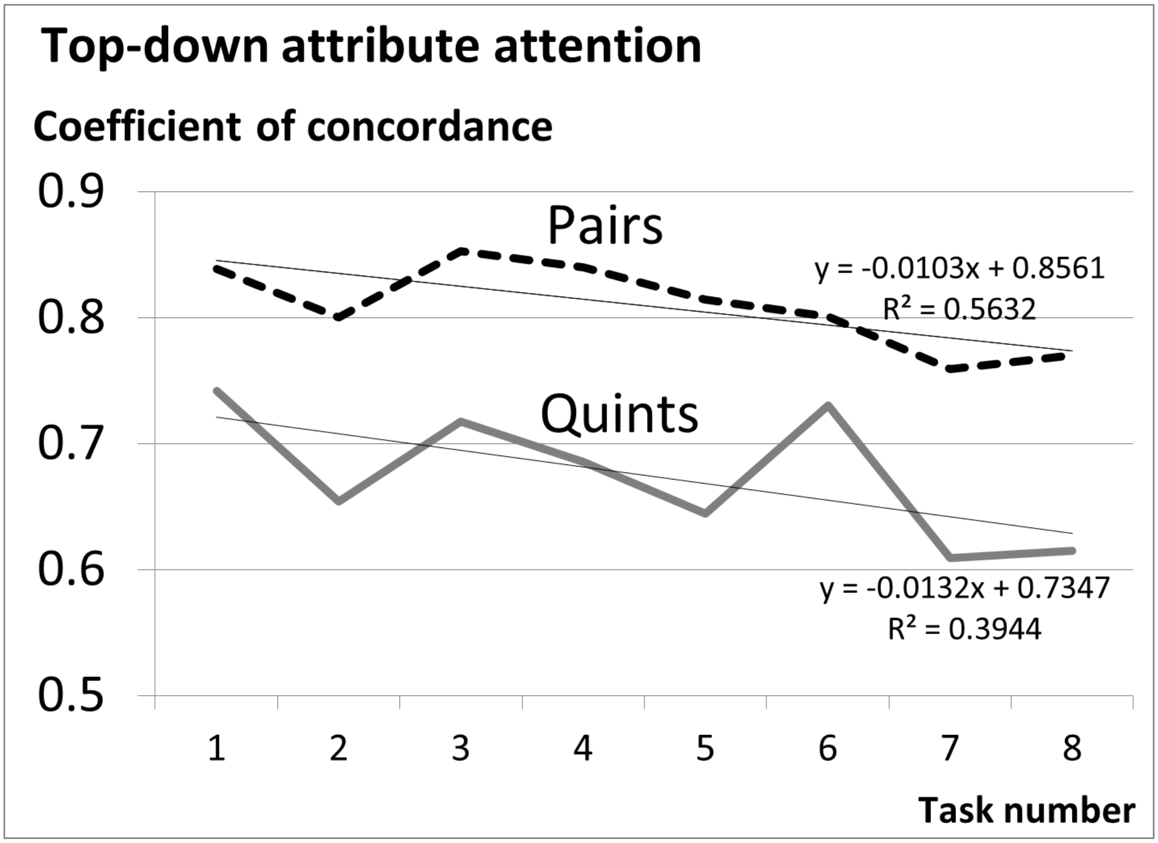
**Figure 6. Search Pattern (Strategy Measure)**



Fourth, we test whether respondents searched the information following a systematic pattern from the top to the bottom of the screen. Because respondents are found to simplify more in quints, they should also be less likely to search information from the top to the bottom of the screen. We therefore defined the following measure to quantify systematic top-down search: We rank all attribute levels with respect to when they were first fixated in a task. The average rank of all levels belonging to an attribute indicates how early the attribute was considered in the search process. We then compare the average ranking of the attributes with a top-down ranking and calculated the coefficient of concordance between the two sets of ranks. A value close to 1 will indicate top-down processing whereas a value close to zero will indicate that attributes are not considered in a schematic way from top to bottom.

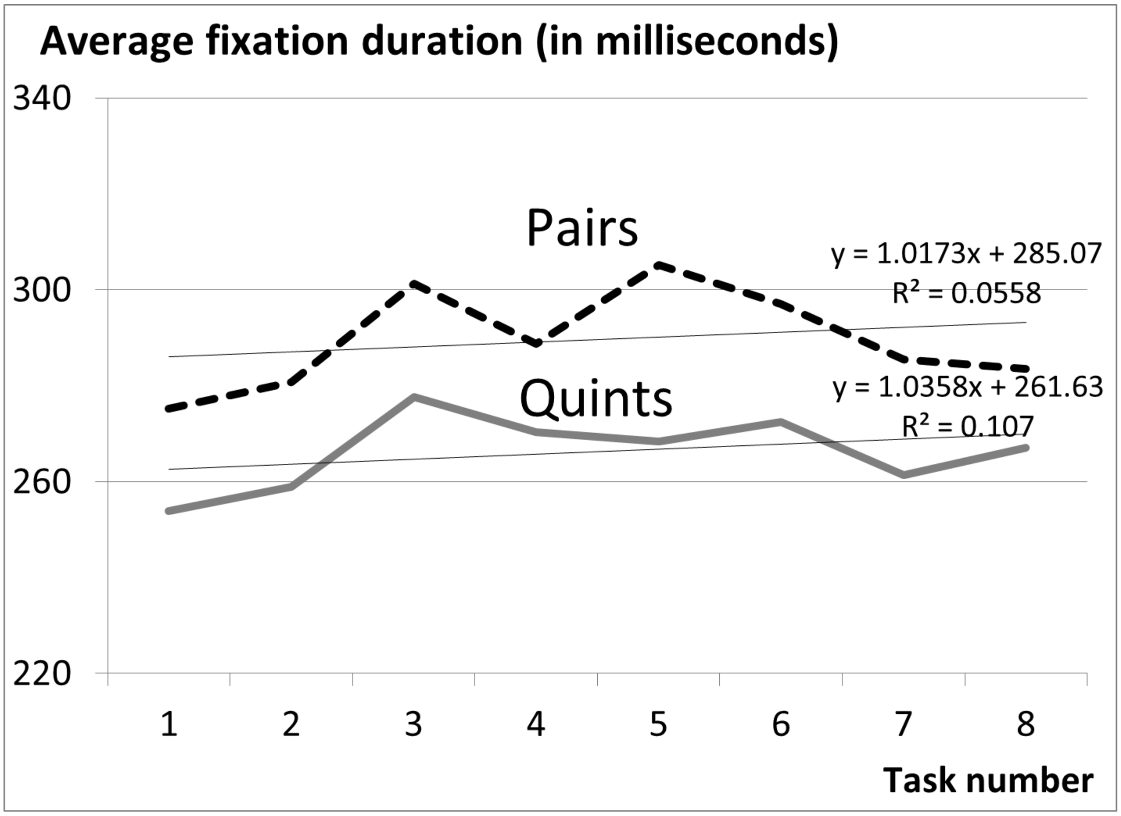
As Figure 7 shows, respondents on average processed the information in the choice tasks more often from the top to the bottom when the tasks included only two instead of five alternatives. This finding is in line with the use of an additive difference strategy in which the respondents compared the alternatives attribute-wise, look at almost all features and process the information from the top to the bottom of the screen. Figure 7 also shows that in later tasks respondents process information less schematically in both the pairs and quints condition.

**Figure 7. Top-Down Attribute Attention**



Fifth, we consider the average fixation duration of all fixations in a choice task. According to the literature very short fixations taking less than 200 milliseconds are often used for scanning and automatic processes, as for example, to understand the structure of a task when the respondent begins processing the information. By contrast, very long fixations might indicate an increased level of processing in more difficult tasks (see e.g., Velichkovsky *et al.* 2002). The average fixation duration for pairs is 296 ms and for quints it is 267 ms. This difference between pairs and quints is statistically significant (t=5.3, p<.01). We interpret this difference in fixation durations as evidence for the use of cognitive processes which involve differencing and adding for pairs, a process that is consistent with the idea that comparisons across alternatives are more time consuming than those within alternatives.

**Figure 8. Average Fixation Duration**



In summary, the average process measures differ strongly between pairs and quints. Pair processing is more thorough, covering proportionately more information, in a more top-down manner, and for greater durations. Quints encourage greater simplification initially as well as over time and lead to deeper processing within a few selected alternatives. Thus pairs fit a model of additive differences while quints reflect a concerted effort to identify a reasonable choice without getting confused by multiple items of available information.

## Task Perception

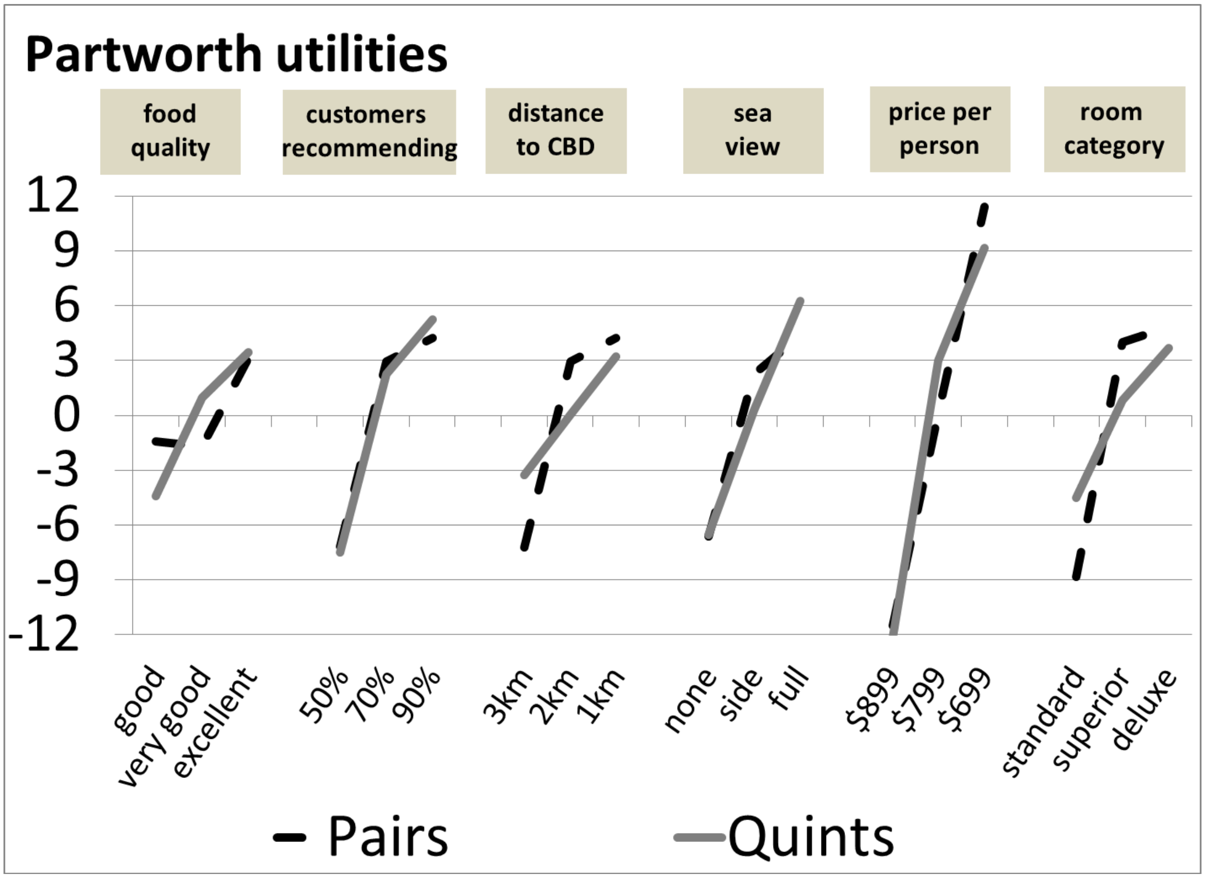
After the initial set of eight choice tasks and before the holdout choice triples, we asked respondents, “How difficult was it for you to choose the vacation package you wanted when last making a choice?” using a 7-point rating scale ranging from “not at all difficult” (-3) to “extremely difficult” (+3). To our surprise, pairs (M=.6, SE=.2) were perceived to be significantly (t=2.9, p<.01) more difficult than quints (M=-.4, SE=.3). Although respondents needed fewer seconds to finish pairs (M=15.0, SE=9.4) compared to quints (M=24.2, SE=14.7), the pairs seem to be cognitively more demanding. Given that most respondents in the pairs condition had to look through most of the information, the tedium of doing that eight times may have made it seem more difficult. By contrast, the goal of finding an acceptable beach vacation was perceived as easier for our respondents in the quints. These differences are also reflected in differences in the patterns of part-worths and predictive accuracy under pairs compared with quints.

## Similarity of the Part-Worths

The part-worth utilities were computed on the individual level by applying Sawtooth Software's Hierarchical Bayes (HB) multinomial logit (MNL) estimation. As shown in Figure 9, the average part-worth utilities are similar, with a correlation of r=.92. Contrary to our expectations, visually it appears that pairs demonstrate greater non-linearity in valuations within attributes. This result suggests that it is more likely that non-linear cutoff values were used when respondents answered the pairs questions.

Next, we also analyzed attribute importance weights. We calculated attribute importances by calculating the ratio of the range of an attribute's utilities against the sum of the ranges across all attributes. The correlation of the average importance weights is also high (r=.86). Importantly, pairs elevate unimportant attributes. The mean of the standard deviation of importances across respondents is 20% less for pairs than for quints (M(pairs)=.097, M(quints)=.117; t=-3.1, p<.01). That finding is consistent with pairs generating a focus on all attributes. The increased attention on less important attributes increases the relative importance of these attributes in the decision process.

**Figure 9. Comparison of the Part-Worth Utilities in Pairs and Quints**



### Predictive Performance

To evaluate the predictive performance, we first looked at the internal hit rates and the output from Sawtooth Software's Hierarchical Bayes estimation. The key measures are included in Table 2.

It does not make much sense to directly compare the internal hit rates for pairs and quints, because the probability of correctly predicting a pair at random is 50%, but is only 20% for quints. The average hit rate is 72% for pairs and therefore is 22% above chance level. For quints, the improvement above chance level is 35%, given a hit rate of 55%.

Because it is hard to correct hit rates for the number of alternatives in the choice set, percent certainty, or another likelihood-based statistic, is a more appropriate indicator of model fit. It is an information-theoretic measure that compares the information explained by the model to the total uncertainty of the system (Hauser 1978). The measures included in Table 2 show that the internal model fit is better for pairs than for quints. The percent certainty for pairs is .90 whereas it is only .69 for quints. One explanation for this result might be that respondents in the pairs condition more consistently applied the same (additive difference) strategy, but in case of quints used all kinds of different decision strategies. As a consequence the internal consistency might be lowered for quints.

**Table 2: Predictive Performance Measures**

| Measure | Pairs | Quints |
|---|---|---|
| Percent Certainty | .90 | .69 |
| RLH: Root Likelihood | .93 | .60 |
| Internal hit rate | 72% | 55% |
| Hit rate from cross-validation | 75% | 53% |
| Hit rate predicting holdout triples | 76% | 57% |

Consistent with the low error as indicated by the Percent Certainty, pairs more consistently predicted the holdout triples shown at the end of the survey. The hit rate for pairs is 87% (69%) in the first (second) holdout task whereas it is only 56% (58%) for quints. This difference is significant for the first holdout task ($\chi^2$=6.4, p=.01), and is directionally consistent for the second holdout task ($\chi^2$=1.1, p=.30). There are two possible reasons why pairs predicted the holdout choices better than quints. First, pairs are more similar to triples than quints, meaning that respondents who have frequently used an additive difference strategy in a sequence of pairs might continue to do so in the consecutive triples. Second, the error around pairs to predict holdouts may simply be sufficiently smaller for quints enabling pairs to overcome their 30% deficit in statistical efficiency.

## CONCLUSIONS

### Summary of Empirical Findings

The important lesson is that the decision making process is very different when choosing between two versus higher multiples of alternatives. Those processing differences lead to different patterns of part-worths and predictive accuracy, and suggest contexts in which either task is more appropriate.

For pairs the pattern of fixations and saccades is consistent with an additive difference strategy. That strategy assesses the relative benefit one attribute at a time and sums those differences across attributes to identify the most preferred option. This within-attribute processing has the advantage of enabling an assessment of each attribute independently from the other attributes. We find that an additive difference strategy leads to greater use of available information, with 92% of the pair information fixated on compared with 69% of the quint information.

The process with more alternatives is quite different. For quints, the need for simplification across 30 pieces of information encourages the use of an important attribute to rule out less promising options. The process of finding a good option from many alternatives can best be described as a search that over time gets more effective at focusing on less information to identify a satisfactory choice. The variability in the search strategy for quints contrasts a relatively mechanistic and stable choice process on pairs.

In terms of efficiency, compared with quints, pairs took 40% less time, were 31% less statistically efficient, but generated 33% more accurate predictions of holdout triples and were more consistent internally. Our results therefore contradict Pinnell and Englert (1997) who find that pairs are no better at predicting holdouts. Perhaps because of the need for accuracy, pairs are perceived to be substantially more difficult. We suggest that this difference can be explained with the cognitive process in pairs which seems to be more demanding.

The average part-worths from the pairs and the quints seem similar, with a correlation of .92. However, differences in details matter. Pairs provide more discrimination of levels within attributes while quints reveal greater discrimination across attributes. In particular, visual inspection suggests greater non-linearity within attributes for pairs, e.g., revealing a large difference between poor and good food compared with good to excellent. This finding is in line with the results by Pinnell and Englert (1997) who observe a “non-linear relationship indicating a loss aversion effect” (p. 150) for pairs. By contrast, for quints the relationship between the three levels is far more linear. That said, quints reveal 20% greater separation in the relative importance of attributes; it appears the complexity of having more attributes reveals respondents’ strategy to focus attention on more important attributes. Here, our results contradict the earlier findings by Pinnell and Englert (1997) who found that “important attributes are more important in pairs” (p. 150).

### How Many Options Should You Use in Your Conjoint Study?

We suggest that pairs are appropriate if one wants efficient measures of how people use many attributes to make choices. Pairs also make sense when the choice is difficult or highly emotional. When patients make choices that involve trading off substantial loss of income and hospital time against longer expected life, having just two options makes the decisions less overwhelming.

However, where the goal is to simulate choices where there are many alternatives and relatively few attributes, then a multi-alternative CBC is appropriate. A good example would be shelf studies that explore consumer ability to find preferred brands in a complex display and respond to different promotional efforts. Put differently, if the decision process involves substantial simplification to find a reasonable option from a large set, then there are advantages to showing a greater number of alternatives per conjoint choice set.

We remain surprised at finding that fewer than 10% of the studies reported in the past four Sawtooth Software Proceedings used pairs. One reason for the lack of use of pairs may stem from the well-known finding that having many alternatives improves the technical statistical efficiency of the design. A second and more reasonable problem with pairs arises from the process revealed by eye-tracking. The additive difference process may be more effective at revealing consistent tradeoffs, but may be even farther removed from what happens in the marketplace with many attributes and many alternatives.

That said, we believe that pairs are underutilized. Apart from greater efficiency, pairs are appropriate when decisions are sufficiently important that simplification to a few attributes makes little normative sense. Furthermore, pairs will be more efficient at assessing consumer reaction to changes in all attributes, in cases where decisions are very important thus justifying consideration of all attributes. Pairs also are reasonable when the attributes are novel, or where when respondents have deep emotional reactions. In the latter situations pairs might help respondents because a weighted additive process facilitates the thoughtful integration of all the attributes of a decision.

Martin Meissner

Harmen Oppewal

Joel Huber

## References

Bettman, J., Johnson, E. J., & Payne, J. W. (1991). Consumer Decision Making. In T. Robertson & H. Kassarjian (Eds.), Handbook of Consumer Behavior (50–84). NJ: Prentice-Hall: Englewood Cliffs.

Böckenholt, U., & Hynan, L. S. (1994). Caveats on a Process-Tracing Measure and a Remedy. Journal of Behavioral Decision Making, 7(2), 103–117.

Bunch, D., Louviere, J., & Anderson, D. (1983). “A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes,” Working Paper, Graduate School of Business, University of California, Davis (available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.196.4913&rep=rep1&type=pdf).

Eggers, F., Hauser, J. R., & Selove, M. (2016). The Effects of Incentive Alignment, Realistic Images, Video Instructions, and Ceteris Paribus Instructions on Willingness to Pay and Price Equilibria. Sawtooth Software Conference Proceedings 2016.

Ford, J. K., Schmitt, N., Schechtman, S. L., Hults, B. M., & Doherty, M. L. (1989). Process Tracing Methods: Contributions, Problems, and Neglected Research Questions. Organizational Behavior and Human Decision Processes, 43(1), 75–117.

Hauser, J. R. (1978). Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. Operations Research, 26(3), 406–421.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press: Oxford.

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), 350–367.

Meissner, M., & Decker, R. (2010). Eye-Tracking Information Processing in Choice-Based Conjoint Analysis. International Journal of Market Research, 52(5), 593–610.

Meissner, M., Musalem, A., & Huber, J. (2016). Eye-Tracking Reveals a Process of Conjoint Choice That Is Quick, Efficient and Largely Free from Contextual Biases. Journal of Marketing Research, 53(1), 1–17.

Orme, B. (2013). The CBC System for Choice-Based Conjoint Analysis—Version 8. Sawtooth Software Inc., Orem, Utah, (accessed October 10, 2016, available at http://www.sawtoothsoftware.com/support/technical-papers/cbc-related-papers/cbc-technical-paper-2013).

Orquin, J. L., Bagger, M. P., & Mueller Loose, S. (2013). Learning Affects Top Down and Bottom Up Modulation of Eye Movements in Decision Making. Judgment and Decision Making, 8(6), 700–716.

Payne, J. W. (1976). Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis. Organizational Behavior and Human Performance, 16(2), 366–387.

Payne, J. W., Bettman, J. R., Coupey, E., & Johnson, E. J. (1992). A Constructive Process View of Decision Making: Multiple Strategies in Judgment and Choice. Acta Psychologica, 80 (1–3), 107–141.

Pinnell, J., & Englert, S. (1997). The Number of Choice Alternatives in Discrete Choice Modeling. Sawtooth Software Conference Proceedings 1997, 121–153, (available at https://www.sawtoothsoftware.com/download/techpap/1997Proceedings.pdf#page=135).

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. Psychological Bulletin, 124(3), 372–422.

Russo, J. E., & Dosher, B. A. (1983). Strategies for Multiattribute Binary Choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9(4), 676–696.

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The Role of Process Data in the Development and Testing of Process Models of Judgment and Decision Making. Judgment and Decision Making, 6(8), 733–739.

Stüttgen, P., Boatwright, P., & Monroe, R. T. (2012). A Satisficing Choice Model. Marketing Science, 31(6), 878–899.

Tobii Software (2016). Tobii Studio 3.4.5 User Manual. (accessed October 10, 2016, available at http://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-studio-user-manual.pdf).

Todd, P. M. (2007). How Much Information Do We Need? European Journal of Operational Research, 177 (3), 1317–1332.

Velichkovsky, B. M., Rothert, A., Kopf, M., Dornhöfer, S. M., & Joos, M. (2002). Towards an Express-Diagnostics for Level of Processing and Hazard Perception. Transportation Research Part F: Traffic Psychology and Behaviour, 5(2), 145–156.

Yang, L., Toubia, O., & De Jong, M. G. (2015). A Bounded Rationality Model of Information Search and Choice in Preference Measurement. Journal of Marketing Research, 52(2), 166–183.

# FINDINGS OF THE 2016 SAWTOOTH SOFTWARE CBC MODELING PRIZE COMPETITION

**BRYAN ORME**
*SAWTOOTH SOFTWARE*

## INTRODUCTION

CBC (Choice-Based Conjoint) is the most widely used conjoint analysis method today. Among Sawtooth Software users, HB estimation for CBC is by far the most common utility estimation approach. Most users stick with CBC/HB software's default settings: main effects estimation (considering only the independent effects of each attribute) with the part-worth utility function (effects-coding). We have seen a long trail of evidence at Sawtooth Software conferences and in the academic literature that these strategies lead to very good models. We have long believed that the resulting market simulators do a superb job predicting the shares of choice for the variety of choice scenarios our clients might specify. Despite our confidence, *we don't want to be complacent*. We designed the 2016 Sawtooth Software CBC Modeling Prize to bring together a diverse group of teams to test those assertions and to see if other models and software might do significantly better. We have always believed that different approaches can be very successful and that no one method consistently dominates, so the diversity of the top-performing submissions in this competition was no surprise to us.

In designing the 2016 CBC Modeling Prize we were inspired by the $1MM Netflix Prize and patterned our approach after it in terms of managing the process and making it a robust test from a statistical perspective. The key element was how to keep the winning team from just overfitting to the holdouts with spurious or nonsensical parameters. The economic reality was that we couldn't offer a $1MM prize, so we consoled the participants that the opportunity to win and present the winning model here at the Sawtooth Software conference was well worth the $995,000 prize gap. Fifteen[1] teams ended up joining the competition. To win the $5,000 prize, the best team needed to beat the default approach (HB main effects) as well as a more sophisticated ensemble approach (involving a combination of 20 HB and 20 latent class models) seeded by Sawtooth Software.

## THE COMPETITION SETUP

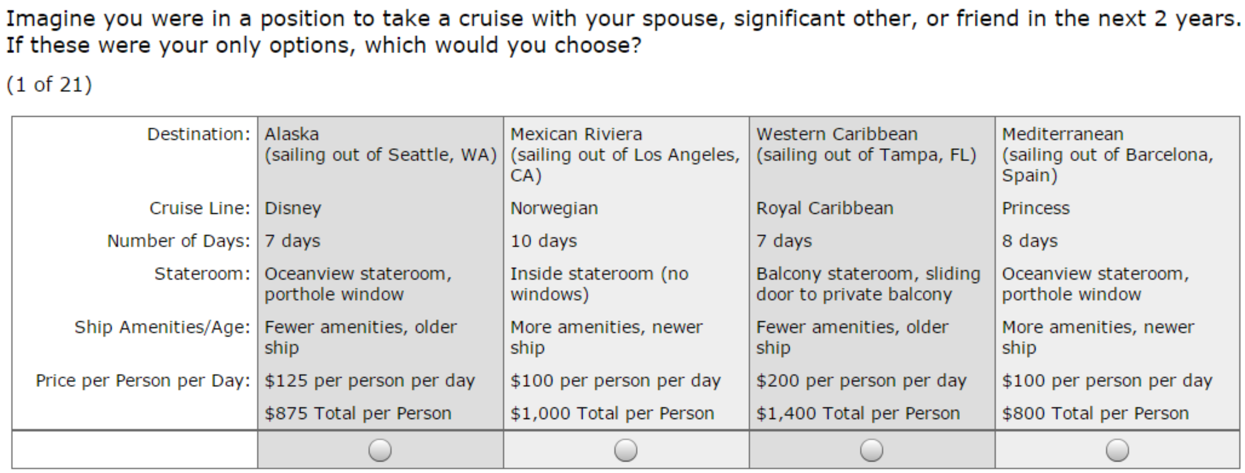
We designed a typical CBC study involving choice of vacation cruise packages on six attributes (in a 6x6x5x3x2x5 design, shown in Appendix A). Around 1350 panelists from SSI completed a 10-minute questionnaire (inter-quartile range 7–14 minutes) including 21 CBC questions along with a few other questions regarding past travel behavior, disposable income for travel, and attitudes/preferences about cruise vacations (including BYO questions regarding preferred levels of each non-ordered conjoint attribute). After cleaning out the fastest and least consistent respondents, we were left with 1200 completed respondent records. When respondents

[1] It is rather striking that only one of the fifteen teams was led by a US-based researcher. Maybe this says something about the lack of extra time researchers in the US have to work on R&D projects like this? Could it be that US teams are more money motivated and $5000 prize money just wasn't enough to capture their interest? It is certainly interesting to think about why only *one* of the teams was based in the US!

entered the survey, we randomly assigned them to one of two buckets: the calibration sample (n=600) or the holdout sample (n=600). The questionnaires looked identical, so respondents did not know they were in one sample or the other.

Each respondent completed 21 CBC tasks with four alternatives per screen (Exhibit 1).

**Exhibit 1. Sample Choice Task**



Imagine you were in a position to take a cruise with your spouse, significant other, or friend in the next 2 years. If these were your only options, which would you choose?

(1 of 21)

| | | | | |
|---|---|---|---|---|
| Destination: | Alaska<br>(sailing out of Seattle, WA) | Mexican Riviera<br>(sailing out of Los Angeles, CA) | Western Caribbean<br>(sailing out of Tampa, FL) | Mediterranean<br>(sailing out of Barcelona, Spain) |
| Cruise Line: | Disney | Norwegian | Royal Caribbean | Princess |
| Number of Days: | 7 days | 10 days | 7 days | 8 days |
| Stateroom: | Oceanview stateroom, porthole window | Inside stateroom (no windows) | Balcony stateroom, sliding door to private balcony | Oceanview stateroom, porthole window |
| Ship Amenities/Age: | Fewer amenities, older ship | More amenities, newer ship | Fewer amenities, older ship | More amenities, newer ship |
| Price per Person per Day: | $125 per person per day | $100 per person per day | $200 per person per day | $100 per person per day |
| | $875 Total per Person | $1,000 Total per Person | $1,400 Total per Person | $800 Total per Person |
| | ○ | ○ | ○ | ○ |

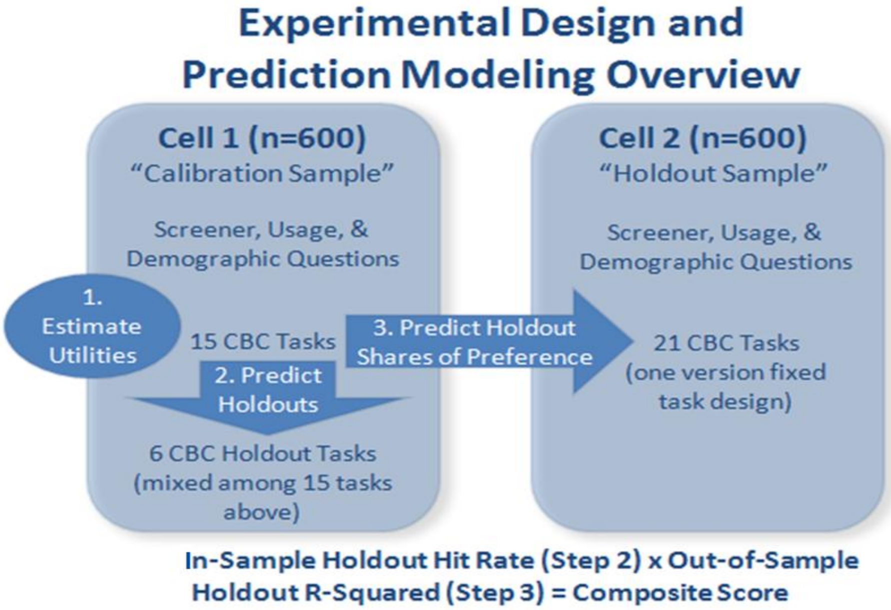
The 600 calibration respondents completed 21 CBC tasks that were experimentally designed using 300 versions (blocks) of Sawtooth Software's balanced overlap[2] design plan. Six of the 21 tasks were holdouts, interspersed throughout the 21 CBC tasks. The holdouts were for predictive validation and not used for utility estimation.

The 600 holdout respondents saw an identical-looking CBC questionnaire where all 21 tasks were fixed across respondents using a single version (block) plan, though we randomized the task presentation order. We purposefully made these 21 holdout tasks difficult to predict: two of the concepts within each task had a great deal of similarity (were defined using the same levels across 3, 4, or 5 of the 6 attributes). Not only were these holdout tasks trickier to predict, but the enhanced similarity between concepts made them actually more like competitive offerings one sees in the real world.

Exhibit 2 is a schematic showing the study design along with the three main steps for submitting predictions.

[2] Balanced Overlap plans are near-perfectly level balanced (both one-way and two-way attribute level occurrence) and nearly orthogonal both within and across versions (blocks). They also feature a modest amount of level overlap (levels repeated across more than one choice) within each task.

**Exhibit 2**



The criterion for winning the competition was the joint predictive validity for the 6 in-sample choice tasks (raw first choice hit rate) and the 21 out-of-sample choice tasks (R-squared based on the share of preference probabilities). Teams could use different models to predict the in-sample and out-of-sample holdouts, though we reserved a special honorable mention category (the "one model wonder") for the single model that did the best job predicting both types of holdouts. Interestingly enough, the grand prize winner (Naji Nassar) was also the "one model wonder," though his model didn't achieve either the absolute best in-sample hit rates or the best out-of-sample share predictions.

We invited teams to join the competition through an open call that was published on our website, in our LinkedIn group, the *Quirk's Marketing Research Review*, and the American Marketing Association's *Marketing Insights* magazine. Fifteen teams of researchers entered the competition. Sawtooth Software also seeded the competition with three submissions. The prize for winning the competition was $5,000 plus a free registration to the 2016 Sawtooth Software Conference with the opportunity to present the winning model at the conference and publish that model within these Sawtooth Software Proceedings. The winners were:

- **1st place:** MIReS: Naji Nassar (Marketing Intelligence & Research Services)
- **2nd place:** Team Nutcracker: Dmitry Belyakov (Ipsos Comcon)
- **3rd place:** SKIM 2 Team: Marco Hoogerbrugge, Kees van der Wagt, Remco Don, and Bingqian Gao (all SKIM Group)

Honorable mentions went to:

- **Best In-Sample Hit Rate:** Landsberger Strasse 2 Team: Merlin Müller, Stefan Binner, Isabella Geisselhardt, Maximilian Rausch, and Markus Böttger (TNS Infratest and bms Marketing Research + Strategy)
- **Best Out-Of-Sample Predictions:** SKIM 2 Team: Marco Hoogerbrugge, Kees van der Wagt, Remco Don, and Binqian Gao (all SKIM Group)

- **One-Model Wonder:** MIReS: Naji Nassar (Marketing Intelligence & Research Services)

We will see below that there was actually very little margin separating the top teams. Furthermore, a variety of statistical approaches and software were very successful.

The competition ran from November 1, 2015 to July 29, 2016. Teams could submit predictions once per week during that period and once per day over the last two weeks of the competition. We leveraged ideas from the Netflix Prize competition to avoid the possibility that the teams would just continue to iterate their solutions to overfit the holdouts. For each submission, week after week, we scored the predictive models (using an automated script) and reported the results on a tracking leaderboard on Sawtooth Software's website. However, for leaderboard tracking we only used a randomly selected half of the holdout tasks (the "quiz" holdouts, randomly selected once at the beginning of the competition and held constant throughout the competition). This allowed the teams to get a good feel for how well their models were doing, but would penalize the teams if they built models that took advantage of variation in just the "quiz" half of the holdouts that wouldn't generalize well to all of the holdouts (the "quiz" + "test" holdouts). At the very end of the competition, the teams were scored based on all the holdouts.

Sawtooth Software seeded the competition with three submissions. The approach and model specifications for these were set prior to seeing the data and (unlike the other teams) Sawtooth Software was not allowed to iterate and try to improve the fit to the holdouts (other than to adjust the scale factor to improve the out-of-sample R-squared to the "quiz" holdouts). These three seed approaches were:

- HB-MNL (using Sawtooth Software's CBC/HB system) under main effects part-worth estimation with default settings (prior variance = 2, degrees of freedom = 5)
- Same as above, but tuned to the calibration tasks for optimal prior variance and prior degrees of freedom[3]
- An ensemble of 20 different latent class and 20 HB solutions (described in Appendix A)

The three seed submissions from Sawtooth Software all performed well, with *ensemble* > *priors optimized HB* > *default HB* in terms of relative predictive validity.

A stipulation for winning the $5,000 prize was that the winner had to beat the best of the Sawtooth Software seed submissions (the *ensemble* solution), which the top five teams were all able to do.

## Inference vs. Prediction

Market researchers and economists often debate the value of inference versus prediction. For conjoint/choice analysis, inference often focuses on interpreting coefficients, such as whether one level of an attribute is preferred to another for the population or how much people are willing to pay for one feature over another. Prediction, however, deals with such issues as what people will choose when given a set of product alternatives defined on multiple attributes. For

[3] The priors were not tuned to the holdouts. Rather, we used Sawtooth Software's CBC/HB Model Explorer program to jackknife across calibration tasks to find the best combination of priors to fit the data. Orme and Williams demonstrated this method of fine-tuning HB priors at the SKIM/Sawtooth Software European Conference in Rome in 2016. Sawtooth Software has released a software tool called the *CBC/HB Model Explorer* for this purpose.

this CBC modeling competition, the competitors had to be ready to predict the mind-bogglingly vast number of choice scenarios that respondents were never asked to consider.

## The Challenge of CBC Prediction

Teams estimated part-worth utilities using the fifteen CBC tasks each respondent answered (for the n=600 Cell 1 respondents), where each task involved a choice among four concepts. The attribute list makes it possible to construct 6x6x5x3x2x5=5400 unique product concepts. Assuming we don't duplicate concepts within the same choice task (which the design didn't), there are 35 trillion possible choice scenarios, assuming order of concepts does not matter[4]. The challenge facing the teams competing in this competition was to build a model that could do a creditable job predicting the choices that people could make for any of those 35 trillion possible situations. Of course, we couldn't actually test their ability to predict all 35 trillion potential tasks accurately. We selected just 21 of them to be evaluated by the holdout (Cell 2) respondents and 300 versions x 6 tasks = 1800 holdout tasks for the calibration respondents (Cell 1) to evaluate.

## Prediction Results

**Exhibit 3. Final Leaderboard Results**

| | | | | Quiz + Test Results (All Holdouts) | | |
|---|---|---|---|---|---|---|
| Rank | Team Name | Submission Date | Single Utility Model[5] | Within-Sample Hit Rate | Out-of-Sample R-Squared | Composite Score |
| 1 | MIReS | 7/29/2016 | Yes | 0.6844 | 0.9129 | 0.6248 |
| 2 | Nutcracker | 7/29/2016 | No | 0.6831 | 0.9123 | 0.6232 |
| 3 | SKIM Team 2 | 7/29/2016 | No | 0.6775 | 0.9130 | 0.6186 |
| 4 | Landsberger Strasse | 6/3/2016 | Yes | 0.6831 | 0.9048 | 0.6181 |
| 5 | Scooter-QX | 6/20/2016 | No | 0.6825 | 0.9043 | 0.6172 |
| 6 | Sawtooth Software3 (Ensemble 20LC&20HB) | 11/1/2015 | Yes | 0.6814 | 0.9056 | 0.6170 |
| 7 | Sawtooth Software2 (Priors Optimized HB) | 11/1/2015 | Yes | 0.6792 | 0.9070 | 0.6160 |
| 8 | Illuminas Partners | 5/17/2016 | Yes | 0.6761 | 0.9099 | 0.6152 |
| 9 | Landsberger Strasse2 | 7/29/2016 | No | 0.6850 | 0.8959 | 0.6137 |
| 10 | Yoda | 7/29/2016 | No | 0.6794 | 0.9021 | 0.6129 |
| 11 | SKIM Team 1 | 7/29/2016 | No | 0.6750 | 0.9062 | 0.6117 |
| 12 | Displayr | 7/26/2016 | Yes | 0.6789 | 0.8991 | 0.6104 |
| 13 | Sawtooth Software Fan Club | 2/24/2016 | Yes | 0.6747 | 0.9016 | 0.6083 |
| 14 | Sawtooth Software1 (Default HB run) | 11/1/2015 | Yes | 0.6725 | 0.9042 | 0.6080 |
| 15 | Jedi Dragons | 7/29/2016 | Yes | 0.6667 | 0.9018 | 0.6012 |
| 16 | Prediction Addiction | 6/24/2016 | Yes | 0.6594 | 0.8981 | 0.5922 |
| 17 | Jigsaw | 3/7/2016 | Yes | 0.6481 | 0.9013 | 0.5841 |
| 18 | SMAP | 7/28/2016 | Yes | 0.6486 | 0.8893 | 0.5768 |

[4] Some of the teams submitted predictions that took concept order into account. If we assume that order matters (predicting choice among concepts A, B, C, D is different from predicting choice among A, B, D, C, etc.), then there are 24 times more possible scenarios to predict, or 849 trillion!

[5] Indicates that the team used a single utility model to predict both the in-sample and out-of-sample holdouts.

It is interesting to note the parity among the top 14 submissions (where the 14th is the default Sawtooth Software submission using CBC/HB software). This baseline submission (Sawtooth Software 1 Default HB run) achieved a composite score of 60.8% and the best submission (MIReS) achieved 62.5%. This seems pretty close, though it is notoriously hard to move the needle much in terms of holdout predictions for conjoint analysis.

As a point of comparison, the winning team for the $1MM Netflix prize bested Netflix's predictions of movie ratings by 10.06% (measured in terms of reduction in RMSE of movie ratings). Team MIReS' out-of-sample share predictions had an RMSE of 3.22, compared to 3.38 for the default Sawtooth Software HB run, a reduction of 4.7% in RMSE. One wonders, then, if a) the default Sawtooth Software modeling approach is closer to theoretical optimal prediction power for CBC than Netflix was with predicting movie ratings of its users, b) Netflix's prediction problem for movie ratings is just harder than CBC predictions, c) we weren't able to attract as deep a pool of world-class modeling talent as Netflix was able to do with its $1MM bounty. We think there's some degree of truth to all three hypotheses, even though we're quite confident that some of the best conjoint modelers in the world entered our CBC modeling competition.

Exhibit 4 gives a very brief summary of the modeling approach used by each of the top ten teams.

**Exhibit 4. Approaches Used by Top 10 Teams**

| Rank | Team | Methods & Model Specification |
|---|---|---|
| 1 | MIReS | HB (GAUSS implementation), interaction between Stateroom and Ship Amenities, utility constraints, budget constraints, fuzzy consideration set model |
| 2 | Nutcracker | Sawtooth Software CBC/HB. 16 separate models in an ensemble, varying priors, covariates, utility constraints, and attribute codings |
| 3 | SKIM Team 2 | Sawtooth Software CBC/HB single model for hit rates with ASC for position effect; regression based ensemble of many different HB solutions for share predictions (i.e., each single HB prediction was treated as one predictor) |
| 4 | Landsberger Strasse | HB (R bayesm, with a Dirichlet Process Prior) |
| 5 | Scooter-QX | Latent Class Analysis (Q Software) ensemble of 20 solutions using a mixture modeling approach, distributional assumption: Multivariate Normal – Full Covariance |
| 6 | Sawtooth Software 3 | HB and Latent Class (Sawtooth Software) ensemble of 40 utility runs with varying starting points for LC and varying covariates for HB, default part-worth coding |
| 7 | Sawtooth Software 2 | Sawtooth Software CBC/HB one priors-optimized run, default part-worth coding |
| 8 | Illuminas Partners | Sawtooth Software CBC/HB single run, default priors, with position-specific effects for order within a task |
| 9 | Landsberger Strasse 2 | Sawtooth Software CBC/HB 4 separate models in an ensemble with different covariates for hit rates; shares of preference based on HB models run separately within 5 LC segments |
| 10 | Yoda | Sawtooth Software CBC/HB, survey questions about preferences for levels of unordered attributes coded as augmented tasks, some utility constraints |

Naji Nassar of team MIReS provides a detailed discussion of his winning model in the next paper in these Proceedings.

Nine of the top 10 submissions used HB estimation, though sometimes with different software implementations and sometimes using different distributional assumptions. Five of the top 10 submissions used an ensemble of models. Four of the top 10 submissions used HB with covariates. We should be careful about drawing too many conclusions from this summary of the top performers since they certainly were biased: influenced by the seed Sawtooth Software submissions, by Sawtooth Software documentation, and past Sawtooth Software conference presentations. There is no doubt a lot of self-selection bias towards HB usage.

Below is an exhaustive list of all the software and utility estimation algorithms used across the 15 teams:

- Sawtooth Software's CBC/HB
- Sawtooth Software's Latent Class, CCEA
- Q
- R (ChoiceModelR, mlogit, bayesm)
- Nlogit
- GAUSS

It's also quite interesting to look at the models that the different teams specified. The list below is not meant to be an exhaustive list of what was attempted, but meant to give an example of the wide variety of strategies.

- Using covariates
- Estimating price as linear, part-worth, or thermometer coding
- Constrained vs. unconstrained utilities
- Using subsets of tasks
- Using respondent choices to other questions in the survey about the attribute levels (BYO questions on certain non-ordered conjoint attributes)
    - As 1) Covariates, 2) Data Augmentation, 3) As individual-level utility constraints
- Ensembles or single models
- Estimating separate models within subsets of respondents
- Modeling alternative number as ASCs (to account for concept order tendencies)
- Examining interaction effects and alt-spec attributes
- Accounting for attribute non-attendance
- Coding budget constraints

## Randomized First Choice vs. Draws

In 1998, prior to the widespread use of HB for conjoint analysis, the author developed a market simulation approach to reduce IIA (Independence from Irrelevant Alternatives, also known as the red-bus/blue-bus problem). Called Randomized First Choice ("RFC"), it is a highly flexible approach since it can be used with part-worth utilities coming from any utility estimation approach. But, if you are using HB and are able to use multiple draws per respondent, using HB draws is more sophisticated and statistically sound. Randomized First Choice could be described as simulating poor man's draws (independent draws with equal variance across all part-worths).

Because of the strength of this dataset for examining the accuracy of market simulations when the scenarios include pairs of highly similar alternatives, we decided to look again at a comparison of simulating on HB draws (respondent-level draws of beta) vs. Randomized First Choice on the point estimates.

We compared the use of the logit (share of preference) rule with the draws to Randomized First Choice operating on the point estimates, tuning the scale factor to best predict the out-of-sample aggregate share predictions for the 21 holdout scenarios. The results are shown in Exhibit 5. For simulating on the draws we used 200 draws per respondent across 600 respondents, for a total of 120,000 utility vectors.

**Exhibit 5. Randomized First Choice vs. Share of Preference Accuracy**

| | **Mean Absolute Error (MAE)** | **R-Squared Fit** |
|---|---|---|
| Share of Preference (logit) on the draws | 2.63 | 0.9092 |
| Randomized First Choice on point estimates | 2.65 | 0.9073 |
| Share of Preference (logit) on the point estimates | 2.72 | 0.9004 |

The three simulation methods (all operating on individual-level HB data) perform quite well. Randomized First Choice works slightly better than Share of Preference on the point estimates. Share of Preference operating on the level of the draws works a tiny bit better—but the results are so similar as to be essentially a tie[6].

If you want the convenience and speed of using Randomized First Choice within Sawtooth Software's market simulation software, RFC works very well. But, simulating on the draws is a more sophisticated and statistically sound way to use HB results. If you can build a market simulator that operates on the level of the draws, you will have built a very good market simulator indeed that will be more defensible in academic circles. These two approaches yield extremely similar results in aggregate: Randomized First Choice shares and aggregated shares from the draws have a correlation of 0.9986 for this dataset.

## The Strength of Ensembles

> *This property—group forecasts beat best individual ones—has been found to be true in almost every field in which it has been studied.*
> *—Nate Silver,* The Signal and the Noise

Ensembles (predictions from different models combined) outperformed single models in the $1MM Netflix Prize contest which focused on improving individual-level predictions of movie ratings based on individuals' ratings of other movies. At the 2015 Sawtooth Software Conference, Kevin Lattery demonstrated that ensembles of high-dimensionality latent class solutions could beat the default HB approach in terms of in-sample individual-level hit rates for

[6] At the request of our reviewer, David Lyon, we also looked at results for applying RFC to the HB draws. The predictive outcome was very good (slightly better than share of preference on the draws on one measure of fit and slightly worse on the other, depending on how we tuned the scale factor).

CBC (Lattery 2015). The author independently confirmed Lattery's findings regarding latent class and also demonstrated that ensembles of different HB solutions (each using different covariates) could also beat the default single HB model for CBC in terms of hit rates (Orme 2015).

Going into this 2016 Sawtooth Software CBC Modeling Prize competition, we again expected that ensembles would outperform individual models in terms of hit rates (individual-level in-sample holdout predictions). *What we didn't know and we believe has never been tested before is whether ensembles could improve out-of-sample accuracy of share predictions for holdout scenarios.* We were excited to test this possibility for CBC.

Across the 15 participating teams plus the 3 seed submissions by Sawtooth Software, we had access to 178 total predictive submissions. The vast majority of these submissions were of very good quality. A few of them involved errors in data processing or model building—they had terrible predictive accuracy and obviously were outliers. We sorted the 178 submissions in terms of their "quiz" hit rates and discarded the worst submissions (we did not look at the "quiz" plus "test" hit rates to prioritize the submissions—that would have given us an unfair advantage over the information that the teams had at the time of the competition). Then, we simply averaged across the better submissions to create a single prediction for each respondent and each holdout task[7]. Could such a simple averaging across the better submissions beat the best prediction that *any one team* had submitted?

Exhibit 6 shows that averaging across submissions indeed beats the best single submission from *any* one team in terms of in-sample hit rates. This shouldn't surprise us. Nate Silver said it usually happens (and he's the current media darling when it comes to prediction). Related specifically to CBC data, Lattery and Orme also presented evidence of this at the 2015 Sawtooth Software Conference. The best single submission among all 178 tries across all teams was a 68.50% in-sample hit rate achieved by team Landsberger Strasse 2 on 29-July. A very good prediction indeed! But, simply averaging across either the top 75% or 95% of all submissions achieves a hit rate of 68.86%. Even averaging across the best single submission from each team (again in terms of the "quiz" holdouts) improves upon the best single submission from any one team.

[7] For example, for a given holdout and a given respondent, if >50% of the models predicted that the respondent would choose concept A instead of B, we took that as the consensus ensemble prediction.

**Exhibit 6. Hit Rates: Best Single Submission Compared to Ensemble**

| | **IS Hit Rates (Quiz + Test)** |
|---|---|
| Landsberger Strasse 2 (29-Jul) (Best overall team submission, even better than MIReS) | 0.6850 |
| Top 95% of all submissions ensemble | <u>0.6886</u> |
| Top 75% of all submissions ensemble | <u>0.6886</u> |
| Top 50% of all submissions ensemble | 0.6881 |
| Top 25% of all submissions ensemble | 0.6881 |
| Top 10% of all submissions ensemble[8] | 0.6847 |
| Best submission from each team ensemble | 0.6881 |

Now we get to the issue that we were very interested in testing. We believe we are the first to demonstrate the superiority of ensembles for out-of-sample share prediction accuracy in CBC and moreover we believe that our evidence is very compelling. Referring to Exhibit 7, the best single submission among all 178 tries across all teams was a 0.9139 R-Squared achieved by team Nutcracker[9] on 21-July (congratulations!). But, simply averaging across either the top 25% or 50% of all submissions (again judged only on the random half of the holdouts, the "quiz" holdouts) achieves an R-Squared share of preference accuracy 0.9163. Even averaging across the best single submission from each team (in terms of the "quiz" holdouts) beats the best single submission from any one team.

[8] At first glance it may seem surprising that the ensemble of top 10% of all submissions (18 different models) doesn't perform as well as broader ensembles that include worse individually performing models. However, it is easily explained because these 18 models were mostly contributed by the same very active and high-performing team. Thus, they lack diversity. The same issue occurs for Exhibit 7 ensemble reporting.

[9] Interestingly enough, team Nutcracker did not realize that this 21-July submission was the best overall submission made by any team. Each submission was scored only on the random half of the holdouts (the "Quiz" holdouts) whereas we are in hindsight now judging all 178 submissions in terms of all the holdouts (the "Quiz" plus "Test" holdouts). Team Nutcracker continued to iterate and submitted what ended up being a slightly worse model when scored using all the holdouts for their final submission.

**Exhibit 7. Share Prediction Accuracy:**
**Best Single Submission Compared to Ensemble**

| | OOS R-Squared (Quiz + Test) |
|---|---|
| Nutcracker (21-July) (Best overall team submission, even better than MIReS) | 0.9139 |
| Top 95% of all submissions ensemble | 0.9151 |
| Top 75% of all submissions ensemble | 0.9157 |
| Top 50% of all submissions ensemble | 0.9163 |
| Top 25% of all submissions ensemble | 0.9163 |
| Top 10% of all submissions ensemble | 0.9129 |
| Best submission from each team ensemble | 0.9158 |

Ensembles benefit from both diversity and quality. Thus, it's very helpful if you can combine models that have been built in quite different ways. (For example, it isn't enough to ensemble a series of HB runs whose only difference is random starting seed.) If a single researcher cannot think creatively enough to develop diverse yet quality solutions, then it can be helpful to use the results of multiple independent-thinking researchers.

One challenge for implementing these findings in the real world is what to do when you don't have such a strong set of out-of-sample holdouts as we had here (which is almost certainly the case) to enable you to discard the worst models and avoid including them in the ensemble. We have a straightforward suggestion: create ten or so randomly generated choice scenarios each with, say, four product concepts. Next, compute shares of preference for these scenarios across the sample. This leads to 10 x 4 = 40 share of preference predictions of individual product concepts that you can compare across the candidate models you are thinking about including in the ensemble. Next, compute a correlation table showing how similar the predictions are between all pairs of candidate models. Summarize the average correlation for each model with every other model, which allows you to sort the models from most similar predictions to the others to least similar. As you examine the rank-order of models in your list, if you detect a sudden and dramatic drop-off in terms of average correlation of predictions relative to the other models, the remaining models are probably outliers representing poor quality solutions and should be discarded from the ensemble.

## SUMMARY OF FINDINGS

So what did we learn from the 2016 Sawtooth Software Prize Competition? Here are our observations:

- The default Sawtooth Software approach of using CBC/HB with main effect, part-worth models works very well (at least for this particular data set). The best models devised by 15 teams of researchers could improve only slightly upon the default standard.

- Tuning HB models in terms of the prior variance and prior degrees of freedom can improve both in-sample and out-of-sample holdout predictions (not cheating by tuning to the holdouts of course, but by jackknifing across calibration tasks for holdout validation).
- HB works very well but is not the only way to obtain excellent predictions for CBC. Many different software systems, utility estimation algorithms, and modeling specifications can be successful. No single approach dominates.
- Ensembles of models that are diverse and of high quality can beat the best single submission made by any one superb researcher.
- Ensembles of models are helpful for lifting both in-sample hit rates *and out-of-sample share of preference prediction accuracy* for CBC (this latter finding has never been demonstrated before).
- Simulating on HB draws works just as well and potentially a tiny bit better than simulating using Randomized First Choice on the point estimates.
- An enthusiastic group of analysts will commit a great deal of effort toward these kinds of competitions, not only because they want to boost our collective knowledge about CBC modeling, but also because they think this kind of activity is rewarding—even *fun.* (Our correspondence encompassing hundreds of emails with the teams confirms this).
- It takes hundreds of hours as the competition organizer to pull off a competition like this! But it was certainly rewarding work.

## CLOSING THOUGHTS

Practitioners work in an environment that usually allows limited time and budget for modeling and simulator building. The work involved in building dozens of diverse models and ensembling them is not very practical given the realities of the marketplace. One wonders about the practical improvements or additional insights clients could gain due to an increase of 1 or 2 absolute points in predictive validity or a reduction of 5% in RMSE. But, when competing in a modeling contest, such gains mean the difference between winning and losing!

Consider a properly tuned HB model (for prior variance and d.f.) operating on the level of the draws versus a more sophisticated ensemble of a dozen or more diverse models. Would a manager make a very different decision based on the modest additional lift in predictive validity? No doubt a slightly improved model could easily lead to changing a feature or two or slightly changing the price, but would the impact on share and profitability be significantly different? Given the amount of money and time invested in CBC modeling (by managers, modelers, and respondents) it makes sense to do more with the data we already have paid for to get better answers—especially when hundreds of thousands or millions of dollars may be on the line in terms of potential profits.

For high-end modelers, ensembles offer a competitive advantage and additional point of differentiation. For software developers, certain kinds of ensembles of models could be automated, such as the sequential use of a variety of different meaningful covariates across replicates for both latent class and HB estimation.

Bryan Orme

## APPENDIX A

### CBC Attribute & Level List

Attribute 1: Destination

1. Mexican Riviera (sailing out of Los Angeles, CA)
2. Eastern Caribbean (sailing out of Fort Lauderdale, FL)
3. Western Caribbean (sailing out of Tampa, FL)
4. Alaska (sailing out of Seattle, WA)
5. Norway and Northern Europe (sailing out of Oslo, Norway)
6. Mediterranean (sailing out of Barcelona, Spain)

Attribute 2: Cruise Line

1. Norwegian
2. Disney
3. Royal Caribbean
4. Princess
5. Holland America
6. Carnival

Attribute 3: Number of Days

1. 7 days
2. 8 days
3. 9 days
4. 10 days
5. 11 days

Attribute 4: Stateroom

1. Inside stateroom (no windows)
2. Oceanview stateroom, porthole window
3. Balcony stateroom, sliding door to private balcony

Attribute 5: Ship Amenities/Age:

1. Fewer amenities, older ship
2. More amenities, newer ship

Attribute 6: Price per Person per Day

1. $100 per person per day
2. $125 per person per day
3. $150 per person per day
4. $175 per person per day
5. $200 per person per day

(Note: total price per person was also displayed below the price per person per day, computed as total days x price per person per day.)

## APPENDIX B

### Description of Sawtooth Software's Benchmark Ensemble Solution

(Inspired by Lattery's 2015 Sawtooth Software Conference presentation.) An ensemble of Latent Class and HB solutions (using simple averaging to obtain consensus), where the ensemble contains 40 replicates:

- 20 replicates of Latent Class 24-group solutions, broken out early such that the last 10 iterations provide about 0.1% total lift in log likelihood. Pseudo individual-level utilities for each replicate developed by taking the weighted average of the part-worth utilities, where the weights are each respondent's probability of membership in each group.
- 20 replicates of HB solutions. First, optimal priors (prior DF and prior variance) were searched on the training data set using jackknife and bootstrap resampling. These optimal priors were used in all HB replicates in the ensemble. Each HB replicate was estimated using alternating sets of covariates developed using combinations of survey questions as covariates.

*Method of predicting individual-level choices for Sawtooth Software's Ensemble Solution*: Shares of preference for each in-sample holdout task to be computed using the logit rule for each of the 40 replicates. The individual hit rates were computed by averaging the shares of preference across the 40 replicates for each respondent, thus determining for each holdout task which concept has the highest share of preference and is the most likely choice for each respondent.

*Method of predicting shares of preference for the OOS fixed holdout tasks for Sawtooth Software's Ensemble Solution*: Randomized First Choice (stacking the raw individual-level utilities for all the replicates in the ensemble, such that the respondent utility run contains nxr cases in the conjoint simulator representing n respondents by r replicates), tuned to the *Quiz* data set for optimal scale factor.

## REFERENCES

Lattery, Kevin (2015), "A Machine Learning Approach to Conjoint Analysis: Boosting and Blending Ensembles," *Proceedings of the 2015 Sawtooth Software Conference*, Orem, UT, pp. 353–370.

Orme, Bryan (2015), "Comment on Lattery's Conjoint Analysis Ensembles," *Proceedings of the 2015 Sawtooth Software Conference*, Orem, UT, pp. 371–378.

Nassar, Naji (2016), “The Winning Choice Model: A Semi-Compensatory One,” *Proceedings of the 2016 Sawtooth Software Conference* (this volume), Orem, UT, next pages.

# THE WINNING CHOICE MODEL: A SEMI-COMPENSATORY ONE

**NAJI NASSAR**
*MIRES*[1]

In this paper, we will present the model developed by MIReS that won the 2016 Sawtooth Software CBC Modeling Prize competition[2]. We will describe the general approach to building a choice model and then explain the MIReS perspective, based on our experience and marketing intelligence strategy. Next we'll elaborate on the hypotheses we built and how those hypotheses led us to create the winning model. The paper will conclude with the discussion of future research.

Our decision to participate in the Sawtooth Software predictive modeling competition was two-fold:

- *As a practitioner*: at the early stage, we wanted to treat the competition as a real case for a marketing manager. This choice meant that we eliminated time consuming approaches, like developing ensembles of solutions, or too new statistical techniques. We wanted to focus on our approach's performance factors.
- *As a competitor*: From our point of view, it was very likely that a good number of competitors would base their approaches on the Multinomial Logit model. It's the most widely used one, and several tools are available to estimate it (Sawtooth Software tools, some R packages, Stata, SAS among others). So, we paid special attention at key steps to see how we could gain a competitive advantage when building our own model.

We didn't develop a holistic approach to test several different models and then deliver the best one. We will show how we focused instead on building the most appropriate model for the given data. Thus, external validity of our approach is somewhat limited, but it was a successful method for creating a winning model. We will demonstrate that building an adequate representation of market dynamics can beat widely used models with advanced estimation techniques.

## THE GENERAL APPROACH

This is the criterion that we received: "To win, the team needs to beat the default approach (HB main effects) as well as a more sophisticated ensemble approach (involving a combination of 20 HB and 20 latent class models) seeded by Sawtooth Software." The objective of the competition was to deliver a model that predicts consumers' choices, both at the individual level (by achieving the best within-sample hit rate), and at the aggregate level (by delivering the best out-of-sample share predictions). We followed the two-step approach described by Leeflang *et al.* (2014):

[1] Marketing Intelligence and Research Services

[2] The design of the competition is detailed in Bryan Orme's article "Findings of the 2016 Sawtooth Software CBC Modeling Prize Competition," the preceding paper in this volume.

## STEP 1: CHOICE MODEL

How does the consumer evaluate the proposed alternatives before choosing the most attractive one? What would be the formal model that describes the process of his or her choices? Is it a compensatory model? Is there some hierarchy in the differentiating characteristics? Is there any aspect that eliminates an alternative from evaluation and consideration?

A modeler can consider several choice models (MNL, Paired Comparison, Nested MNL, constrained MNL, MNP, and all their extensions; see Manrai 1995 and Garrow 2009 for some non-exhaustive extensions) and test their performance to represent consumer/respondent choice behavior.

MIReS had an existing baseline model: the compensatory multinomial logit model. We decided to investigate whether there is some departure from this widely used model. Several competitors used such a model, so we were aware that we would have to build the very best choice model possible to differentiate us.

## STEP 2: HETEROGENEITY

To build a choice model for 600 respondents, we needed a representation of consumer heterogeneity over the choice model. Two issues must be considered when describing consumer heterogeneity: its scope and nature.

**Exhibit 1. Heterogeneity representation**

| | | **Nature** | |
|---|---|---|---|
| **Scope** | Discrete | Continuous | Mixed+ |
| Choice model | | [redacted] | [redacted] |
| Consideration set | | | [redacted] |
| Attributes' preferences | | | |

### Choice Model Heterogeneity

After the selection of potential choice models, it can be determined either that:

- Consumers' behavior can be represented by one of those choice models (homogeneity), or
- Consumers can be allocated to several classes, and each class has its own formal representation of choice behavior (discrete heterogeneity).

Our approach was the widely used one of assuming consumer homogeneity in the choice model, i.e., the same model applies for all respondents.

### Consideration Set Heterogeneity

There are numerous publications that examine the consideration set (such as Shocker *et al.* 1991, Horowitz and Louviere 1995, Andrews and Srinivasan 1995). Some publications have treated consumer choice as a two-step process; the first looks at the formation of the consideration set, the second step consists of choosing an alternative from the consideration set. The authors just cited used the Crisp Set Approach, where an alternative is assumed to be either considered or not. Other publications saw operational consideration sets as indicators of preferences, a kind of elimination by cutoff. Other authors used the Fuzzy Set Approach (Bronnenberg and Vanhonacker 1996, Wu and Rangaswamy 2003) where each alternative has a probability to be considered.

Despite those differences, all publications agreed that consideration sets can provide information about preferences that can increase the efficiency of a choice model. One can consider one situation among the following:

- Respondents are considering all the alternatives present at each choice situation.
- Each respondent has his own fixed consideration set (when data have been collected for such exercise, or when, for example, the modeler decides to eliminate the worst alternative from consideration set).
- During a choice situation, every possible set of proposed alternatives has its own probability to be the respondent's consideration set.
- Every alternative has its own probability to belong to respondent's consideration set.

## Attribute Preference Heterogeneity

Twenty years ago, Carroll and Green (1995) stated, "New developments in conjoint analysis are arriving so fast that even specialists find it difficult to keep up. Hierarchical Bayes models, latent class choice modeling and individualized hybrid models are only a few of the approaches and techniques that are arriving on the research scene." No doubt this sentence is still accurate. A complete review of this issue would require an exhaustive approach that goes beyond the scope of this article. However, to represent consumers' heterogeneity across alternative evaluation and attributes perceptions, the modeler can choose among:

- Discrete representation of consumer heterogeneity: Latent Class approach,
- Continuous representation of consumer heterogeneity: Hierarchical Bayes approach (see Andrews, Ainslie and Currim 2002 or Hess, Ben-Akiva and Gopinath 2011 for a comparison between both representations),
- Mixed approach of previous representations and empirical distribution: Augmented Latent Class (Varki and Chintagunta 2004), or Mixture of Distributions (Train 2008, Train 2016) and many others.

## Estimation

Once the choice model has been designed and heterogeneity defined, one can estimate the model using various statistical techniques. Three of those techniques have been developed in the last decade or two:

| | **Advantages** | **Disadvantages** |
|---|---|---|
| Hierarchical Bayes | Easy to implement, excellent predictive capacity, very fast, by far the most common utility estimation approach. Most importantly, our learning curve in such approach has already been climbed | Only normal distribution (and mathematical transformations of it) can be practically used to describe consumer heterogeneity. |

| | | |
|---|---|---|
| Simulated Maximum Likelihood | All previous hypotheses (as to choice model, heterogeneity presentation, etc.) can be estimated | An optimization procedure that is computationally cumbersome |
| Expectation Maximization | Easiest one to implement, even for Mixture of Distributions | Limited to normal distributions, time consuming computations |

These steps are interactive:

- Coherence must be achieved between the choice model and the heterogeneity representation.
- Statistical estimation must be able to handle the models and heterogeneity assumptions made by the modeler. If not, (s)he must change either the model or the estimation technique.

## PRIOR BELIEFS AND CHOICES

We describe the general approach we follow to build a decision model at the individual level. For the Sawtooth Software competition, we wanted to replicate real case conditions in terms of delays and time expended. We did not experiment with all the modeling opportunities just mentioned, but took some important decisions up-front in our approach:

- We decided to build a unique model that would describe the marketing phenomena of the marketplace. In our mind, the aggregate market was the aggregation of individual behavior and we felt it was most important to deliver one, and only one, accurate description of market dynamics to the marketing manager. So, we did not use separate models for in-sample and out-of-sample prediction.
- Respondents' decisions would be captured by one choice model, with the starting point being the MNL model where an alternative's utility is explained by the 7 attributes detailed in Appendix A (6 attributes of the design of experiments, plus total price per person).

$$\begin{aligned} Util(Cruise) \\ &= \delta_{Destination} + \delta_{CruiseLine} \\ &+ \delta_{NumberofDays} + \delta_{StateRoom} + \delta_{ShipAmenities} - \delta_{Price} - \delta_{Budget} \end{aligned}$$

- Consumer heterogeneity in terms of attributes' part-worths was considered to be continuous and able to be captured by normal distribution.
- Hierarchical Bayes estimation was the appropriate statistical approach given our objectives (easy to implement, established predictive validity).

We chose to focus our modeling efforts on:

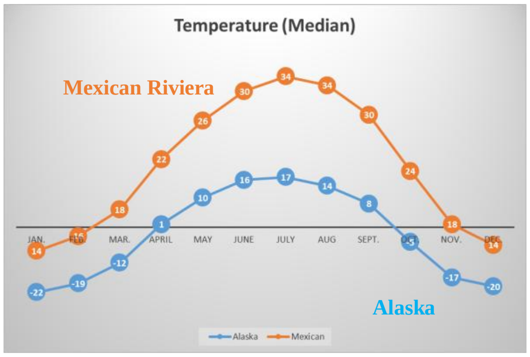
- Choice model formulation: relationship between the 7 attributes and alternatives' utilities, and
- Consideration set model: which attributes had an impact on consideration set formation, and how to interact the consideration model with the choice model.

## Our Choice Model

We had no prior information on the consumer process, which was how (s)he chose to take a cruise. To overcome this, we conducted some qualitative interviews to investigate how consumers choose a cruise. We also ran some counting and quantitative analyses. We structured those around the relevant attributes, and used them to generate the structure of our choice model.

### Destination

Some destinations seemed to be eliminated from consideration sets. For example, "I'll never buy a cruise with Alaska as the destination; it's too cold" was a typical qualitative comment. However, the underlying reason was unobserved from other attributes or survey data.



We determined that destinations can have some physical/perceived aspects which eliminate the associated cruises from consideration set, no matter the levels of other attributes of the cruise.

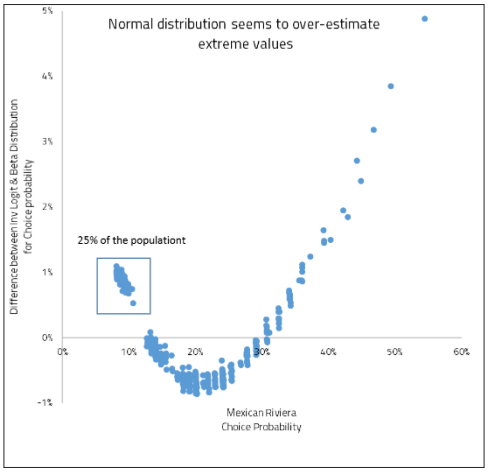
Counting analysis confirmed to us that the normal distribution doesn't seem to be the most accurate distribution to represent respondents' heterogeneity in terms of destination preferences. With the data having been collected using Sawtooth Software's Balanced Overlap design plan, all destinations are more or less equally proposed to each respondent ($N_d$ between 9 and 11 times each over the 15 choice situations). We counted the number of times the respondent chose each destination ($n_d$) then made the following hypothesis:

- At the individual level, the number of times a destination was chosen ($n_d$) follows a binomial distribution with parameter the number of exposures ($N_d$) and choice probability ($p_d$).[3] $n_d \sim Binomial(N_d, p_d)$.
- The choice probability $p_d$ can follow, across respondents, either:

| **Beta distribution** | **Inverse logit of Normal distribution** |
|---|---|
| $p_d \sim Beta(\alpha, \beta)$ | $Logit(p_d) \leftarrow \gamma_d$<br>$\gamma_d \sim Normal(M_\gamma, \sigma_\gamma)$ |
| LogLikelihood: **-2575.8** | LogLikelihood: -2595.9 |

The Beta distribution appeared to represent consumer preferences across destination preferences better than the Normal distribution. We then investigated the discrepancies between both distributions.

[3] We will use Winbugs notation as we used this software for this analysis.



When we compared the distributions of (transformed) Normal draws and Beta draws, we saw that the Normal distribution is underestimating the frequency of low values. For example, with the Mexican Riviera destination, 25% of the population is impacted, so it’s quite important. If we want to correct such a discrepancy, we must add a type of penalty for destination preference (based on Normal distribution).

Conclusion: We will suppose that each destination has a probability to be considered:

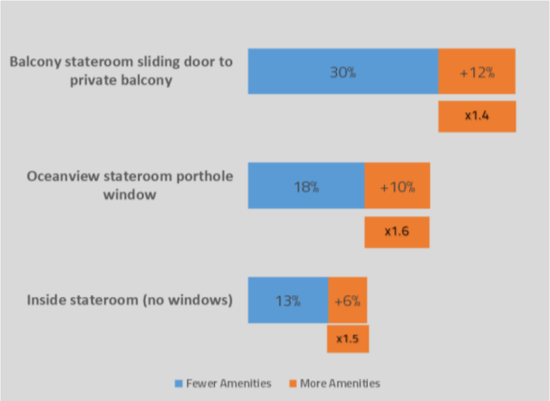
$$\boldsymbol{P_{Destination}} = \boldsymbol{logit^{-1}}(\boldsymbol{\gamma_{Destination}})^{4}$$

By incorporating a consideration probability, we can model the large number of low-utility destination values while keeping a normal prior in the choice model itself, making estimation simpler.

## State Room & Ship Amenities

Counting analysis highlighted an interaction between those two attributes at the aggregate level.

Choice probability (counting)



We tested such interaction by considering it in the preference model. We introduced some logical constraints on this interaction part-worth:

- More Amenities should have higher preferences,
- Oceanview stateroom and Balcony stateroom should be preferred to Inside stateroom.

The cruise utility for the MNL model incorporated this interaction into a single term, becoming:

$$\boldsymbol{Util(Cruise)} = \ \boldsymbol{\delta_{Destination}} + \boldsymbol{\delta_{CruiseLine}} + \boldsymbol{\delta_{NumberofDays}} + \ \boldsymbol{\delta_{StateRoom.ShipAmenities}} - \boldsymbol{\delta_{Price}} - \boldsymbol{\delta_{Budget}}$$

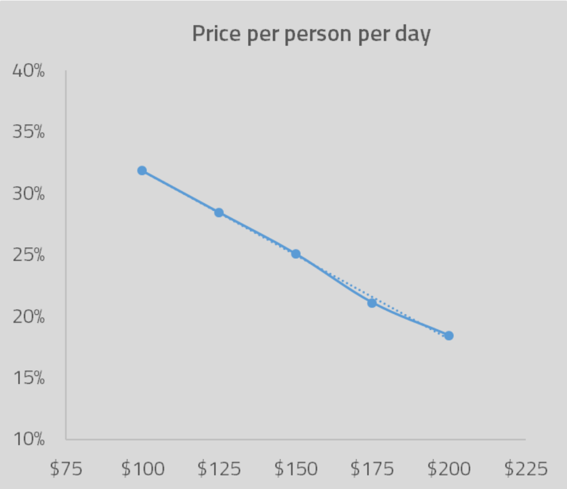
[4] $logit^{-1}(x) = \frac{e^{x}}{1+e^{x}}$

We kept some logical constraints for the part-worth of each level's utility in this interaction term (arrows go from one level to another that must have a higher utility than the starting one)

| **StateRoom Ship Amenities:** | **Fewer** | **More** |
|---|---|---|
| Inside stateroom (no windows) | ↓ ↓ | → |
| Ocean view stateroom porthole window | ↓ ↓ | → |
| Balcony stateroom sliding door to private balcony | ↓ | → |

## Price Per Day

Choice probability (counting)



For the price attribute, we introduced some logical constraints for decreasing part-worth:

$$\mathbf{0 = -\delta_{\$100} < -\delta_{\$125} < -\delta_{\$150} < -\delta_{\$175} < -\delta_{\$200}}$$[5]

## Budget Per Person[6]

We assumed that if the cost of a cruise exceeds the respondent's budget, that cruise won't be considered, no matter the levels of other attributes. ("Budget" in the formula below refers to the total cost per person per day of a cruise.)

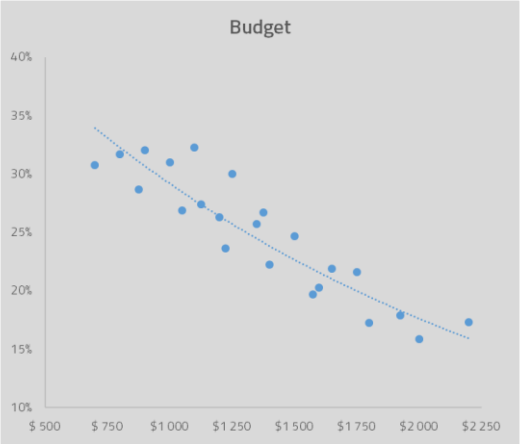
$$P_{Budget} = P(Budget < Budget_{Max}) = logit^{-1}(\gamma_{Budget,1} - \gamma_{Budget,2} * Budget)$$

Consideration must be decreasing with budget, so we imposed a constraint: $-\gamma_{Budget,2} < 0$

[5] Setting the $100 part-worth to zero allows us to generate positive part-worths for all other levels of the attribute.

[6] Computed as the product of number of days and price per person per day, this attribute had more than 20 distinct values. It was treated as a continuous attribute.

Choice probability (counting)



As one can notice, the shape of the relationship between budget and aggregate choice probability is non-linear. So we introduced a second term for budget attribute: its natural logarithm. And we constrained both parameters, one for budget, and one for its logarithm, to be negative at the individual level.

$$-\boldsymbol{\delta_{budget}} < 0$$
$$-\boldsymbol{\delta_{ln(budget)}} < 0$$

The final form of cruise's utility for the MNL model became:

$$\boldsymbol{Util(Cruise)} = \boldsymbol{\delta_{Destination}} + \boldsymbol{\delta_{CruiseLine}} + \boldsymbol{\delta_{NumberofDays}} + \boldsymbol{\delta_{StateRoom.ShipAmenities}} - \boldsymbol{\delta_{Price}} - \boldsymbol{\delta_{budget}Budget} - \boldsymbol{\delta_{ln(budget)}ln\ (Budget)}$$

and we determined that consideration of a cruise is based upon its destination and budget:

$$\boldsymbol{P_{Destination}} = \boldsymbol{logit^{-1}(\gamma_{Destination})}$$
$$\boldsymbol{P_{Budget}} = \boldsymbol{logit^{-1}(\gamma_{Budget,1} - \gamma_{Budget,2} * Budget)}$$

Both conditions must be satisfied for consideration.

In our earlier discussion of general approaches, we cited two ways to treat the consideration set:

- **Crisp Set Approach**: Here an alternative is assumed to be either considered or not. One has to estimate the probability of each possible consideration set. The respondent has to make a choice; the consideration set can't be empty, so in our case, 15 = $(2^4 - 1)$ consideration sets are possible. The probability for each consideration set can be written as[7]:

$$P(\{y_1^4\}) = \sum_{\{y_1^4\}}^{0,1} \frac{\prod\left[(P_{Destination}.\,P_{Budget})^{yi}(1 - P_{Destination}.\,P_{Budget})^{1-yi}\right]}{1 - \prod[(1 - P_{Destination}.\,P_{Budget})]}$$

  This approach supposes that consideration of each cruise is independent from the 3 others in the choice situation. But cruises can share the same destination (this occurred in about 10% of the choice tasks in this design). Further, if a cruise satisfies the budget condition, all cruises that have a lower budget must satisfy it as well. So, independence can't be assumed in our case. One would have to build a specific approach for each possible consideration set. So, we ruled out the crisp set approach.

[7] In this formula, *yi* is 1 if alternative *i* is present in possible consideration set $\{y_1^4\}$; 0 otherwise. The products are over the 4 alternatives in each consideration set.

- **Fuzzy Set Approach**: In this approach, an alternative has a probability to be considered One needs only the marginal distributions of consideration, one for each alternative included in the universal set. We used this approach in our final model:

$$\boldsymbol{Prob(Cruise)} = \frac{\boldsymbol{P_{Destination}.P_{Budget}.\exp\,(Util_{Cruise})}}{\boldsymbol{\sum P_{Destination}.P_{Budget}.\exp\,(Util_i)}}$$

This leads us to a semi-compensatory choice model, as the compensatory term (MNL) is counterbalanced by a non-compensatory term (consideration model).

## Estimation & Simulations

We used the Hierarchical Bayes algorithm to estimate attributes' part-worths and the parameters that figured in our formal model. We followed, point-by-point, every step Kenneth Train (2003) described in his section 12.6, "Hierarchical Bayes for Mixed Logit." We used:

- Starting values: Negative uniform for individual draws (10x), and large covariance matrix (10x the covariance matrix of the uniform draws)
- For the first 10,000 draws, we kept draws showing better likelihoods
- 500,000 draws for convergence
- We retained afterwards 2,500 draws, skipping 100 between two retained draws (so systematically sampled from a series of 250,000 draws)
- We didn't use tying techniques for constrained parameters, instead we transformed normal draws (see Appendix B for more detail).

Since our model reduced the IIA problems, we directly used simulations based on 2,500 HB draws. We then took for each respondent the mean of alternative probabilities over all those draws. The first choice gave us the most likely choice the respondent will make for the 6 in-sample choice tasks (hit rate). The means over respondents of the individual alternatives' probabilities gave us the performance of each alternative for the 21 out-of-sample choice tasks (R-squared based on the share of preference probabilities).

We focused all of our attention on building the choice model itself so we didn't investigate all the avenues of research we mentioned in the general approach. With one model, we were able to establish an adequate description of market dynamics. Our description offers the marketing manager some significant advantages:

- the predictive capacity of the model,
- the non-compensatory role of destination and budget in the choice model, which can have a significant impact on pricing cues,
- market competition as a consideration model defined the most competing alternative, and reduced the IIA problem.

## Limitations and Future Research

Several hypotheses led us to our model based on our experience and descriptive analysis. But we are aware that other hypotheses can't be rejected regarding the consideration model, and our prior beliefs may have biased us when we developed the model.

First of all, we had no information about consideration set formation, even from the survey. We developed our formation from the 15 choices made by respondents. One can't use consideration set formation apart from this choice model. We don't have the in-depth information that consideration formation could provide for the marketing manager: given the destination and budget, the marketing manager might be able to define the addressed market and the main competitors. The main drivers of consideration formation could also be helpful for cruise promotion.

Second, the consideration set can't be empty at any purchase occasion, whatever the characteristics of the proposed alternatives. This can lead to overestimation of the consideration model parameters.

For destination consideration, we mustn't forget that departure from Normal distribution, as the consumers' heterogeneity representation, had been also highlighted. Consideration set parameters can be seen as a "correction" for this departure. We did not have the information for the previous destinations for the respondents who had taken cruises before (47% of the respondents). Variety-seeking buying behavior can also have an impact on the consideration set for this last set of respondents. So, precautions must be taken for destination consideration.

For budget consideration, this factor varied from $700 to $2200 and indicated more than 20 levels. That's quite a wide range and budget was the attribute with the widest variation (more than 50% of the choice tasks showed a budget range[8] higher than $500). There is extensive literature on the impact of an attribute's number of levels and range on attribute importance. Such bias could explain a part of the non-compensatory effect of budget.

The consideration model allowed us to build a semi-compensatory model that captured marketplace dynamics and showed (as we now know) high predictive validity. With the limitations we described above, one can't separate the two parts of our modeling approach.

We referenced some ideas in the general approach that we would like to test further:

- Choice Model: MNL, Nested MNL over destination (limited by the modest overlap), MNP as one can suspect that destinations can show unobserved similarities (weather, for example)
- Consideration Model: Crisp Set approach versus Fuzzy Set approach
- Heterogeneity: Normal versus Dirichlet/Gamma distribution (as extensions of Beta distributions) versus empirical distribution

[8] Measured as difference between lowest and highest budget across proposed cruises.

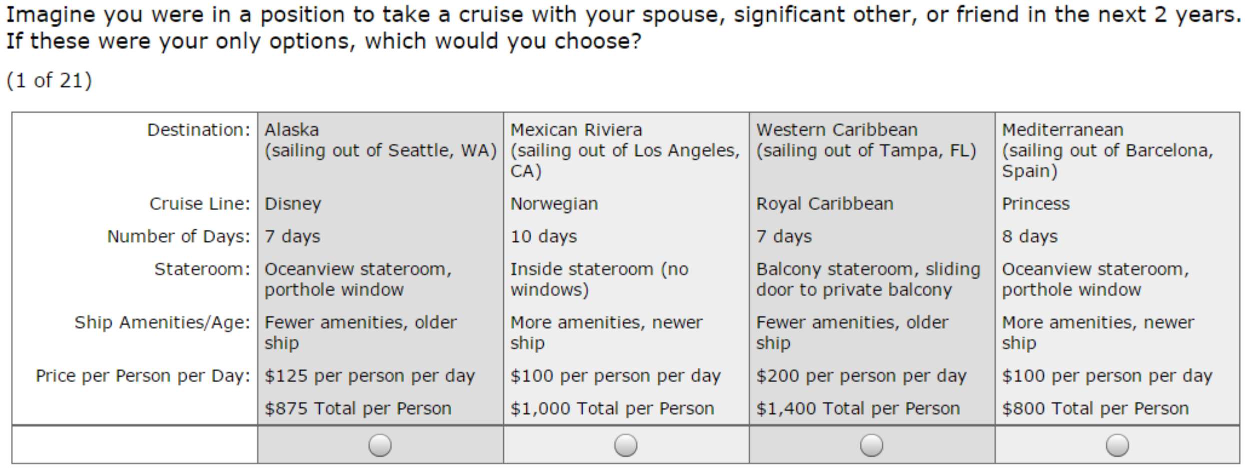
We put relatively minimal effort into building our semi-compensatory choice model, using a level of effort that would be feasible and reasonable in real commercial studies. It could be interesting to see the performance results of our research ideas, and the relative effort necessary to invest in order to achieve additional performance.

Naji Nassar

# Appendix A

## CBC Choice Screen, Attributes & Levels

Each respondent completed 21 CBC tasks with four alternatives per screen, like the following.



Imagine you were in a position to take a cruise with your spouse, significant other, or friend in the next 2 years. If these were your only options, which would you choose?

(1 of 21)

| | | | | |
|---|---|---|---|---|
| Destination: | Alaska<br>(sailing out of Seattle, WA) | Mexican Riviera<br>(sailing out of Los Angeles, CA) | Western Caribbean<br>(sailing out of Tampa, FL) | Mediterranean<br>(sailing out of Barcelona, Spain) |
| Cruise Line: | Disney | Norwegian | Royal Caribbean | Princess |
| Number of Days: | 7 days | 10 days | 7 days | 8 days |
| Stateroom: | Oceanview stateroom, porthole window | Inside stateroom (no windows) | Balcony stateroom, sliding door to private balcony | Oceanview stateroom, porthole window |
| Ship Amenities/Age: | Fewer amenities, older ship | More amenities, newer ship | Fewer amenities, older ship | More amenities, newer ship |
| Price per Person per Day: | $125 per person per day | $100 per person per day | $200 per person per day | $100 per person per day |
| | $875 Total per Person | $1,000 Total per Person | $1,400 Total per Person | $800 Total per Person |
| | ○ | ○ | ○ | ○ |

| **Attribute 1: Destination** | **Attribute 2: Cruise Line** |
|---|---|
| 1. Mexican Riviera<br>(sailing out of Los Angeles, CA)<br>2. Eastern Caribbean<br>(sailing out of Fort Lauderdale, FL)<br>3. Western Caribbean<br>(sailing out of Tampa, FL)<br>4. Alaska<br>(sailing out of Seattle, WA)<br>5. Norway and Northern Europe<br>(sailing out of Oslo, Norway)<br>6. Mediterranean<br>(sailing out of Barcelona, Spain) | 1. Norwegian<br>2. Disney<br>3. Royal Caribbean<br>4. Princess<br>5. Holland America<br>6. Carnival |
| **Attribute 3: Number of Days** | **Attribute 4: Stateroom** |
| 1. 7 days<br>2. 8 days<br>3. 9 days<br>4. 10 days<br>5. 11 days | 1. Inside stateroom (no windows)<br>2. Ocean view stateroom, porthole window<br>3. Balcony stateroom, sliding door to private balcony |
| **Attribute 5: Ship Amenities/Age:** | **Attribute 6: Price per Person per Day** |
| 1. Fewer amenities, older ship<br>2. More amenities, newer ship | 1. $100 per person per day<br>2. $125 per person per day<br>3. $150 per person per day<br>4. $175 per person per day<br>5. $200 per person per day |
| **“Attribute” 7: Price per Person**<br>Computed as number of days x price per person per day | |

# Appendix B

## From Normal Draws ($\boldsymbol{\beta}$) to Model Coefficients ($\delta$ $and$ $\gamma$)

The $\boldsymbol{\beta}$ parameters assumed to have normal priors in the Hierarchical Bayes estimation were transformed to yield final model parameters ($\boldsymbol{\delta}$ for the choice model, $\boldsymbol{\gamma}$ for the consideration model) with the desired constraints. The transformations made constraint violations impossible, no matter what values the $\boldsymbol{\beta}$'s took on.

The final form of a cruise's utility for the MNL model:

$$\boldsymbol{Util(Cruise) = \delta_{Destination} + \delta_{CruiseLine} + \delta_{NumberofDays} + \delta_{StateRoom.ShipAmenities} - \delta_{Price} - \delta_{budget}Budget - \delta_{ln(budget)}ln\ (Budget)}$$

Consideration model:

$$\boldsymbol{P_{Destination} = logit^{-1}(\gamma_{Destination})}$$
$$\boldsymbol{P_{Budget} = logit^{-1}(\gamma_{Budget,1} - \gamma_{Budget,2} * Budget)}$$

Overall probability: $\boldsymbol{Prob(Cruise) = \frac{P_{Destination}.P_{Budget}.\mathrm{exp}\ (Util_{Cruise})}{\sum P_{Destination}.P_{Budget}.\mathrm{exp}\ (Util_i)}}$

| Attribute Level | Transformation from $\boldsymbol{\beta}$ to $\boldsymbol{\delta}$ or $\boldsymbol{\gamma}$ |
|---|---|
| Mexican Riviera | 0: reference |
| Eastern Caribbean | No constraints, $\boldsymbol{\delta_{Destination} = \beta_{1..5}}$ |
| Western Caribbean | |
| Alaska | |
| Norway and Northern Europe | |
| Mediterranean | |
| Norwegian | 0: reference |
| Disney | No constraints, $\boldsymbol{\delta_{CruiseLine} = \beta_{6..10}}$ |
| Royal Caribbean | |
| Princess | |
| Holland America | |
| Carnival | |
| 7 days | 0: reference |
| 8 days | No constraints $\boldsymbol{\delta_{NumberofDays} = \beta_{11..14}}$ |
| 9 days | |
| 10 days | |
| 11 days | |
| Inside stateroom Fewer | $\boldsymbol{\delta_{StateRoom1.ShipAmenities1} = 0}$ : Reference |
| Inside stateroom More | $\boldsymbol{\delta_{StateRoom1.ShipAmenities2} = 20 * e^{\beta_{15}}/(1 + e^{\beta_{15}})}$ |
| Ocean view stateroom Fewer | $\boldsymbol{\delta_{StateRoom2.ShipAmenities1} = \delta_{StateRoom2.ShipAmenities2} * e^{\beta_{16}}/(1 + e^{\beta_{16}})}$ |
| Ocean view stateroom More | $\boldsymbol{\delta_{StateRoom2.ShipAmenities2} = \delta_{StateRoom1.ShipAmenities2} + 20 * e^{\beta_{17}}/(1 + e^{\beta_{17}})}$ |
| Balcony stateroom Fewer | $\boldsymbol{\delta_{StateRoom3.ShipAmenities1} = \delta_{StateRoom3.ShipAmenities2} *}$ |

| | |
|---|---|
| | $e^{\beta_{18}}/\left(1+e^{\beta_{18}}\right)$ |
| Balcony stateroom More | $\delta_{StateRoom2.ShipAmenities2} = \delta_{StateRoom1.ShipAmenities2} + 20 * e^{\beta_{19}}/\left(1+e^{\beta_{19}}\right)$ |
| $100 per person per day | 0: reference |
| $125 per person per day | $\delta_{Price2} = 20 * e^{\beta_{20}}/\left(1+e^{\beta_{20}}\right)$ |
| $150 per person per day | $\delta_{Price3} = \delta_{Price2} + 20 * e^{\beta_{21}}/\left(1+e^{\beta_{21}}\right)$ |
| $175 per person per day | $\delta_{Price4} = \delta_{Price3} + 20 * e^{\beta_{22}}/\left(1+e^{\beta_{22}}\right)$ |
| $200 per person per day | $\delta_{Price5} = \delta_{Price4} + 20 * e^{\beta_{23}}/\left(1+e^{\beta_{23}}\right)$ |
| Budget | $\delta_{budget} = 20 * e^{\beta_{24}}/\left(1+e^{\beta_{24}}\right)$ |
| Ln(Budget) | $\delta_{ln(budget)} = 20 * e^{\beta_{25}}/\left(1+e^{\beta_{25}}\right)$ |
| Mexican Riviera | No constraints, $\gamma_{Destination} = \beta_{26..31}$ |
| Eastern Caribbean | |
| Western Caribbean | |
| Alaska | |
| Norway and Northern Europe | |
| Mediterranean | |
| Consideration Budget | No constraint: $\gamma_{Budget,1} = \beta_{32}$ |
| Budget | $\gamma_{Budget,2} = 20 * e^{\beta_{33}}/\left(1+e^{\beta_{33}}\right)$ |

## REFERENCES

Andrews, R. and TC Srinivasan (1995), "Studying Consideration Effects in Empirical Choice Models Using Scanner Panel Data," *Journal of Marketing Research* (February, 1995), pp. 30–41.

Andrews, R., A. Ainslie and I. Currim (2002), "An Empirical Comparison of Logit Choice Models with Discrete versus Continuous Representations of Heterogeneity," *Journal of Marketing Research* (November, 2002), pp. 479–487.

Bronnenberg, Bart J. and Wilfried R. Vanhonacker (1996), "Limited Choice Sets, Local Price Response and Implied Measures of Price Competition," *Journal of Marketing Research* (May, 1996), pp. 163–173.

Carroll, JD and P. Green (1995), "Psychometric Methods in Marketing Research: Part I, Conjoint Analysis," *Journal of Marketing Research* (November, 1995), pp. 385–391.

Garrow, Laurie (2009), Discrete Choice Modelling and Air Travel Demand, Ashgate Publishing.

Hess, S., M. Ben-Akiva, D. Gopinath and J. Walker (2011), "Advantages of Latent Class over Continuous Mixture of Logit Models," Working Paper, Institute of Transport Studies, University of Leeds.

Horowitz, J. and J. Louviere (1995), "What Is the Role of Consideration Sets in Choice Modelling?" *International Journal of Research in Marketing* (May, 1995), pp. 39–54.

Leeflang, P., J. Wieringa, T. Bijmolt & K. Pauwels (2014), *Modeling Market: Analyzing Marketing Phenomena and Improving Marketing Decision Making*, International Series in Quantitative Marketing, Springer.

Manrai, Ajay (1995), “Mathematical Models of Brand Choice Behavior,” *European Journal of Operational Research* (February, 1995), pp. 1–17.

Shocker, A. D., M. Ben Akiva, B. Boccara and P. Negungadi (1991), “Consideration Set Influences on Consumer Decision-Making and Choice: Issues, Models and Suggestions.” *Marketing Letters* (August, 1991), pp. 181–197.

Train, Kenneth (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press. Note particularly Section 12.6 at pages 302–308.

Train, Kenneth (2008), “EM Algorithms for Nonparametric Estimation of Mixing Distributions,” *Journal of Choice Modelling*, vol. 1, pp. 40–69.

Train, Kenneth (2016), “Mixed Logit with a Flexible Mixing Distribution,” *Journal of Choice Modelling*, vol. 19, pp. 40–53.

Varki, S. and P. Chintagunta (2004), “The Augmented Latent Class Model: Incorporating Additional Heterogeneity in the Latent Class Model for Panel Data,” *Journal of Marketing Research* (May, 2004), pp. 226–233.

Wu, J., and A. Rangaswamy (2003), “Fuzzy Set Model of Search and Consideration with an Application to an Online Market,” *Marketing Science* (Summer, 2003), pp. 411–434.

# USING BAYES' THEOREM TO ADJUST SIMULATED PREFERENCE SHARES TO MARKET REALITY

**DAVID BAKKEN**
*FORESEEABLE FUTURES GROUP*

## ABSTRACT

As every practitioner working with conjoint analysis knows, even with the best possible model, simulated preference shares often do not map to actual market shares. This paper introduces a "new" method of post-estimation adjustment using Bayes' Theorem and compares adjusted results to other post-estimation adjustment methods. External measurements of market performance, previous survey data, or even subjective beliefs may serve as prior beliefs in the application of Bayes' Theorem to adjusting shares predicted from Choice-Based Conjoint studies. The method is applied to two case studies. Results indicate that the method may be appropriate in the face of ignorance or uncertainty about the underlying differences between the predicted and reference market shares.

## INTRODUCTION

Conjoint analysis has proved to be a valuable tool for understanding consumer choices among competing alternatives that can be defined (and differentiated) by observable characteristics ("features" or "attributes") and, in many cases, price. Conjoint analysis derives much of its value from the fact that a set of model parameters estimated from data collected from survey respondents can be incorporated into a "market simulator" that permits "what-if" scenario analyses. Most often such analyses are used to refine a product profile, optimize a portfolio, or model competitive dynamics (such as price competition).

As every practitioner working with conjoint analysis knows, even with the best possible model simulated preference shares often do not map to actual market shares. Discrepancies between model-based simulated preference shares and actual market shares can be due to any number of factors that have been identified elsewhere (e.g., Allenby *et al.* 2005; Orme & Johnson 2006).

Practitioners have employed various techniques for closing the gap between simulated preference shares and in-market results. In general, these techniques rely on either improving the quality of the data and parameter estimates (Allenby *et al.* 2005) or use of various post-estimation adjustments such as multiplicative weighting of predicted preferences shares to achieve a target distribution of shares. Orme and Johnson (2006) have described and evaluated several of these adjustments. In this paper we introduce a method of post-estimation adjustment using Bayes' Theorem. The rationale for this method is that, before we collect data and estimate a model, our best guess (prior belief) for the market shares of the various alternatives is the current observed share. The current observed shares provide information about unobserved factors that may be contributing to the differences between predicted and actual market shares.

For example, actual shares are subject to a number of constraints, such as supply capacity, that are not directly observed but that are reflected in the actual market shares.

## ARE DISCREPANCIES REALLY A PROBLEM?

The discrepancies between simulated and actual shares can be substantial. In one of the case studies described later in this paper, the predicted preference share for one alternative was larger than the actual market share by a factor of nine. In the other study, the predicted preference share for one alternative was five times larger than the actual market share.

From one perspective—that of *internal validity*—differences between model predictions and the real world may not matter much. Consider the case where a company must decide on the set of features that will define a new product. As long as the conjoint model informs the company as to the optimal set of features based on some relevant criterion, the model results can be meaningful and useful even if they do not predict real world market shares. However, this view is shortsighted in many respects. Imagine that in this case the predicted market share is two times the (unobserved) actual share the product will achieve. Actual shares translate to actual volumes, and volume usually determines both revenue and profit. If the economic breakeven point for the product is somewhere between the predicted share and the realized actual share, the firm is likely to lose money by launching the product.

Discrepancies between predicted and actual shares also make it difficult to validate our models and modeling approaches. If we cannot demonstrate that our predicted results at least correlate with real world results, managers may justly question the credibility of our findings and recommendations.

Finally, managers may possess knowledge and hold assumptions about their markets that are not made explicit in the discrete choice model. If results predicted by the model are at odds with that knowledge and those assumptions, they again are likely to challenge the findings and recommendations. Just such a case provided the impetus for testing the Bayesian post-estimation adjustment described in this paper.

## CAUSES OF DISCREPANCIES BETWEEN SIMULATED AND ACTUAL MARKET SHARES

Because the reasons for differences between simulated and actual market shares have been discussed in detail by other authors (Allenby *et al.* 2005, Orme and Johnson 2006), I will only summarize those reasons here. Several of these factors impact the estimated parameter values; our primary interest lies in the impact of violating the assumptions incorporated into choice simulators.

Here are the principal causes of discrepancies between simulated and actual market shares:

- Study design factors
- Data collection errors
- Respondent reliability
- Respondent validity
- Modeling errors
- Violation of simulation assumptions

*Study design factors* include poor or incorrect operationalization of attributes and levels, mismatch between the conjoint method used (e.g., CBC, ACBC, MBC) and the real-world choice architecture, inadequate experimental designs, and poor "pre-conditioning" of respondents with respect to their understanding of the choice context and tasks. Because the experimental choice context is, by necessity, an abstraction or simplification of the real marketplace, we may omit or ignore one or more explanatory variables, resulting in a mis-specified model.

*Data collection errors* include inadequate sample size, improper sampling frame, inaccurate measurement of actual share, and asynchronicity between the survey data and actual market share measurement.

*Respondent reliability* is compromised when respondents answer inconsistently or when their responses change systematically over the course of the choice experiment. In addition to these within-subjects effects, stochastic heterogeneity in traits that may be related to preferences may introduce between-subjects noise. For models using hierarchical Bayesian estimation methods, within-subjects noise should have greater impact on the lower level model, while between-subjects noise should have greater impact on the upper level model.

*Respondent validity* suffers when respondents do not answer realistically. They may choose to answer unrealistically—an example would be consistently choosing higher-priced options when they would not do so in real life—or they may be unable to respond realistically, as when they do not have an adequate understanding of some of the attributes that define the choices. Respondents may be insufficiently motivated to respond accurately. Several researchers have experimented with incentivized conjoint methods to reduce the likelihood of unrealistic responses.

*Modeling errors* include failing to account for interaction effects (for example, failing to include a price cross-effect when modeling choices between bundled and *a la carte* options) and possible mismatch between the model assumptions and the data-generating process (for example, assuming a compensatory, additive process when consumers actually employ screening rules or elimination by aspects).

*Violations of simulator assumptions* include all the differences between the conditions encoded into the simulator and the real world. The choices that generate the model parameters for the simulator typically take place under conditions of 100% awareness of the alternatives and an assumption of 100% availability for each of the alternatives (with a corollary assumption that there are no differential supply constraints for the alternatives). Respondents often are better informed and have greater understanding of the differences between the alternatives than might occur in the marketplace. Simulated shares are instantaneous; all options are assumed to have reached maturity (or to require equal time to reach maturity). Other assumptions typically include that the consumers represented in the simulator have no budget constraints, and that the sales efforts of the different brands are equally effective (a corollary of the 100% awareness assumption).

## What Do We Want from an Adjustment Method?

Any method that we apply to reducing simulator vs. real world discrepancies should satisfy a few important criteria.

First, the method should have some plausible link to the reasons for the differences between simulated and real market shares. To understand why this is important, consider a simple aggregate share adjustment. We might run a simulation set up to reflect the current market (brands, prices and attributes), compare the simulated shares to actual shares and then calculate an adjustment factor for each brand by dividing the predicted shares by the actual shares. Then, for every scenario we simulate, we multiply the results for each brand by that brand's adjustment weight (and then renormalize the shares to 100%). Because this method does not take any explanatory variables into account, we have no way of validating simulations for scenarios other than the current market. On the other hand, if our method explicitly incorporates explanatory variables such as awareness or distribution, we can systematically vary our assumptions about those explanatory variables to test the robustness of our adjustment method.

Second, the method should be *transparent.* Underlying relative changes in preference shares (that is, before external adjustment) between scenarios should be preserved in the post-adjustment results. Orme and Johnson (2006) demonstrate how this is not the case for the simple aggregate adjustment described in the preceding paragraph. That method can produce "reversal anomalies" where the post-adjustment simulated shares actually move in different directions from the un-adjusted shares.

Third, the results of the adjustment should be *predictable.* We should be able to tell from the method whether simulated shares will increase or decrease as a result of the adjustment. If we increase the assumed awareness level for a brand, for example, we expect that brand's simulated share to increase. If we decrease distribution, we expect a brand's simulated share to decrease.

Finally, the method should be *robust*, producing consistent results across the range of input values. Basically, this means that for extreme values of adjustment inputs, such as very low levels of awareness, the adjusted results are not wildly off-base.

## Some Possible Solutions

Although it is something of an inversion to talk of the real world "violating" the simulator assumptions, most approaches to dealing with these violations involve interposing some type of correction to bring the simulator assumptions more in line with the real world. For example, if we know the actual awareness levels of each of the brands in our model, or the actual geographic distribution, we can apply some type of weighting scheme that adjusts the probabilities of choosing each of the brands based on actual awareness or distribution (or both). This type of "external" adjustment (i.e., post-estimation adjustment) depends on our ability to discover or make an educated guess about the parameter values for awareness, distribution, or some other real world factor that differs from the simulator assumption.

Some factors, such as awareness and distribution, have an aggregate effect on market shares. Other factors, such as budget constraints, operate at the level of the individual consumer. While individual-specific factors should be applied to respondent level data, aggregate factors such as awareness can be applied at either the aggregate level (applying the weights to the alternatives) or individual-level (stochastically deciding whether each respondent becomes aware of each alternative, in proportion to the target levels of awareness). Aggregate adjustment has a consistent effect across all respondents. Assigning respondent-level probabilities can lead to different results when the simulator is based on individual-level utilities and therefore the specific mechanism used to adjust at the individual level is an important consideration.

Aggregate external adjustments are fairly easy to construct and apply. By way of comparison, we can imagine building a simulator that explicitly incorporates variables that differ between the simulator and the real world. Bakken (2006) has illustrated the use of agent-based modeling to simulate awareness, consideration set formation, and consumer choice. Given a set of individual decision models derived from Choice-Based Conjoint, this approach overlays a set of stochastic processes to model market choices. Each respondent is treated as an autonomous buyer agent subject to a variety of environmental stimuli, such as advertising and word of mouth. Such models can also incorporate seller (brand) agents that respond to the buyer agents' choices (by changing prices, for example, or increasing the resources devoted to creating awareness). Building this type of expanded simulator requires that we both understand and can parameterize all of these various components of the market system of interest.

In most cases, simpler external adjustment methods will be adequate for closing the gap between predicted and actual shares. Agent-based simulations, however, may be a better choice when we need to model temporal dynamics, such as the rate of adoption and diffusion of a new product over time, or the impact of positive and negative word of mouth.

Orme and Johnson (2006) describe and evaluate several external effects adjustments. These include a simple adjustment for awareness wherein respondents are asked to indicate their awareness of each brand in the model (prior to the choice exercise!) and then "post-processing" the part-worth utilities, setting any part-worth for a brand the respondent is unaware of to an arbitrarily low value so that the respondent's probability of selecting that brand is close to zero.[9]

The same type of adjustment can be applied to unequal distribution if individual level data reflecting access (Orme and Johnson suggest the respondent's zip code or region) to the different alternatives is available. The stochastic micro-simulation method described above for awareness can also be used for unequal distribution in those cases where we do not have information about which respondents will have access to the different choices in the simulation.

Orme and Johnson also describe a method of individual-level utility adjustment that applies a brand-specific constant correction to each respondent's part-worth for the brand. They find these brand-specific correction factors with a simple iterative procedure that employs the ratios between target shares and unadjusted simulated shares.

Orme and Johnson describe an additional approach that calculates respondent-level weights. This method might be appropriate in situations where the discrepancy between predicted and actual shares is due to the fact that the sampled respondents differ from the target population in some unobserved way that interacts with their preferences.

## USING BAYES' THEOREM FOR EXTERNAL ADJUSTMENT OF SIMULATED SHARES

These methods for adjusting simulated preference shares are all based on the assumption that the actual market shares represent "truth" and our goal is to move the simulated shares closer to the actual shares. We can also view the "problem" as arising from two different sources of information about consumer preferences, namely the data provided by our survey respondents

[9] As Orme and Johnson note, this reflects a "micro-level correction." In the absence of self-reported awareness for each brand, we could assign respondents randomly to states of aware or non-aware for each brand with a probability equal to awareness measured in some other way, such as a brand tracker and then adjust the brand part-worths. We would need to run many simulations, repeating the random assignment in each iteration, and then average across the iterations to obtain awareness-adjusted preference shares.

and data on actual market shares, and we can view our objective as the integration of these different information sources. Bayes' Theorem (Bayes 1763) offers a way to combine information from the survey data with other, external information about the market.

While many market research practitioners will be familiar with Bayes' Theorem, it may be helpful to review the calculations using a classic example described by Piatelli-Palmarini (1994) and known as the "Juror's Fallacy." In this example, a taxi driver is accused of a hit and run accident. An eyewitness claims to have seen a blue taxi strike a pedestrian and drive away. The police, understandably, look towards the blue cab company for suspects. We'll assume that, for whatever reason, such as company records of which drivers were on duty, a single plausible suspect is identified. Additional information emerges during the course of the trial: there are only two taxi companies in town and one uses only blue cars and the other only green cars. On the night of the accident, 85% of the taxis on the road were green and 15% were blue. The prosecutor has been diligent and conducted an experiment to determine the reliability of the eyewitness. The experiment revealed that, under conditions identical to those on the night of the accident, the eyewitness was able to correctly identify the color of a taxi 80% of the time. The jurors are charged with determining whether, in fact, the accused was involved in the hit and run accident. Our concern at the moment is determining the probability that the cab seen by the eyewitness was blue (which is analogous to determining the probability that the eyewitness is telling the truth).

This example is called the juror's *fallacy* because our intuitive reaction is likely to be that the probability of a blue cab is the same as the reliability of the eyewitness, or 80%. Even if we feel that the answer cannot be quite that simple, we are likely to assume that it is more probable that the cab was blue than that it was green. If, given the fact of a blue taxi, other evidence points to the accused, the preponderance of the evidence likely would support conviction. However, we have not considered the eyewitness's identification must be *conditioned* on the likelihood that any taxi out that night was blue.

There are a few different ways of writing Bayes' Theorem, but one of the easiest to grasp is the formula used by Silver (2012):

$$Prob(event) = \frac{xy}{\left(xy + z(1-x)\right)}$$

The "event" of interest is the taxi being blue. For the other terms:

$x$ = our "prior probability" that the taxi is blue or our best guess *prior* to the eyewitness appearing. In this case, $x$ = 15%, the proportion of taxis on the road that night that were blue.

$y$ = the probability of the eyewitness's testimony if the taxi is in fact blue (that is, the accuracy of the eyewitness), which = 80%.

$z$ = the probability of the witness saying the taxi is blue if in fact it is green, which = 20%

Dropping these values into the equation gives us:

$$Prob(blue\ taxi) = \frac{0.15 * 0.8}{[(0.15 * 0.8) + (0.2 * 0.85)]} = 0.41$$

So now, instead of our intuitive 80% probability that a blue taxi was involved in the hit and run, our *posterior* probability is only 41%, and the driver of the blue taxi appears much less likely to be guilty. So why does the probability that the driver is guilty decline? It's because the probability that any taxi seen by the witness that night was blue is much lower than the probability that any taxi seen was green (15% versus 85%). Of course, there might be other evidence that would alter our prior probability. Perhaps the two taxi companies tended to operate in different parts of the city. In that case, we would need to refine (or "update") our original guesses and recalculate the probability that the witness saw a blue taxi.

How do we apply this calculation to the problem of "adjusting" simulated preference shares? In other words, how do we set values for **x**, **y** and **z** in our formula?

"X" is our prior belief, or our best guess about the expected market share in the absence of the data from our conjoint study. In the project that prompted me to begin exploring Bayes' Theorem as a possible method for adjusting simulated shares, the client firm's market share was 12%. The study was designed to explore pricing for the next generation of the product that included some feature enhancements. In the absence of the market research, our best guess might be that the new version of the product will sell about the same number of units per year as the current version. For the moment considering only aggregate adjustment, we can set **x** to equal current market share for the product.

"Y" is our hypothesis about the probability that a customer will choose a particular alternative. Again considering just aggregate adjustment, we can set **y** equal to the predicted preference share for the alternative.

Setting a value for **z** may present the greatest challenge. Unlike our blue taxi/green taxi example where we know the error rate for the witness's color identification, we have to think carefully about the way in which we establish the probability that we would observe consumers choosing the target alternative if our "hypothesis"—the simulated preference share—is wrong. In the absence of any other information, we can set **z** equal to the random probability of choosing an alternative, which is 1/(number of alternatives). Thus, if we have five alternatives in our simulation, **z** would be set to 1/5 or 0.2.

Going back to the example above:

**x = 0.12** (current market share)
**y = 0.25** (predicted market share)
**z = 0.2** (based on five alternatives in the simulator)
**P = 0.14** (adjusted share prediction)

Changing our assumption about **z** can have a fairly big impact. If our simulation has eight options and we set **z** equal to 0.125 (meaning that it's less likely that the product would be selected by chance), our adjusted share jumps to 21%.

So far we have looked at applying Bayes' Theorem to adjust aggregate preference shares. If we have individual-level prior beliefs we can apply this adjustment at the individual level. For example, if we are simulating automotive choices and we have captured each respondent's current vehicle make and we know something about the retention rate for each brand, we can calculate Bayes' Theorem separately for each respondent and then average across respondents. To keep things simple, assume that there are only two makes of automobile; 45% of the sample currently own Make A, which has a retention rate of 60%. Make B has a 55% share of current owners and a 70% retention rate. For respondents who currently own Make A, we would set **x** for Make A equal to 0.6 (the retention rate, which is our prior belief about the probability that they will choose Make A). Consider two different respondents who both own Make A. For one respondent, the simulated probability of choosing Make A is 30% before adjustment; for the second respondent, that simulated probability is 70%. Setting **x** to 60% for each individual (with **z** = 50% since there are two alternatives), our *posterior* probability of choosing Make A for the first respondent is now 47% while the posterior probability for the second respondent is 67%.[10]

## Case Studies

This Bayesian method for externally adjusting simulated shares was applied to two market studies. The first study focused on preferences for sliding patio doors. The second study focused on individual choices of health insurance coverage.

### Patio Door Study

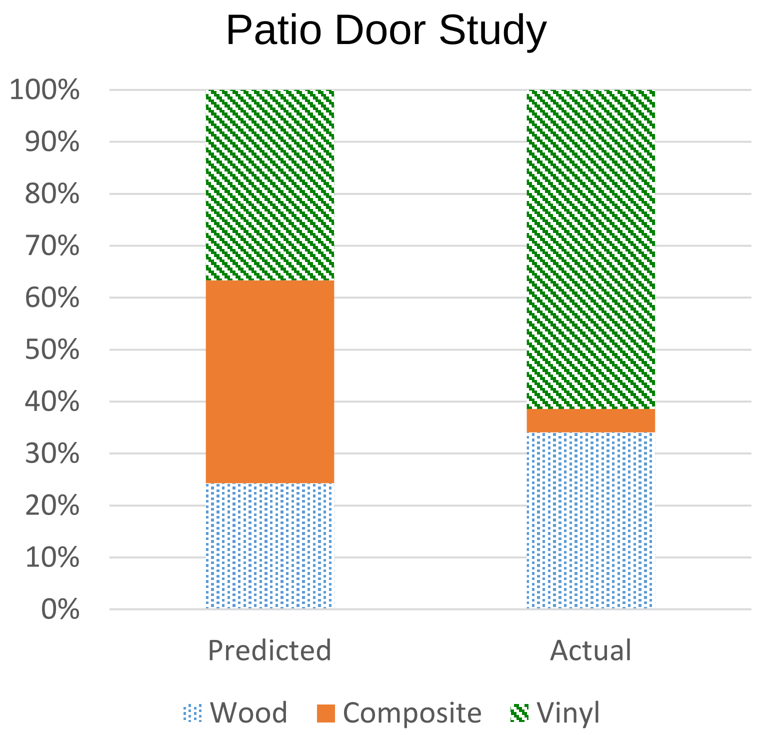
The patio door study was designed to estimate consumers' willingness to pay for specific feature enhancements. Most patio doors are sold through either big box home improvement stores or trade channels (e.g., lumber supply companies and local door and window fabricators who sell direct to contractors). Doors can be constructed using wood (which is usually covered with aluminum cladding on the exterior side), fiberglass/composites, or vinyl. Vinyl doors are the least expensive and account for the lion's share of category volume. Fiberglass/composite doors are more expensive and more durable, but represent only about 3–5% of total patio door unit sales. Wood doors tend to be most expensive and represent about 30–35% of the market. Reasons for the differences across the different materials include awareness, distribution, and supply constraints.

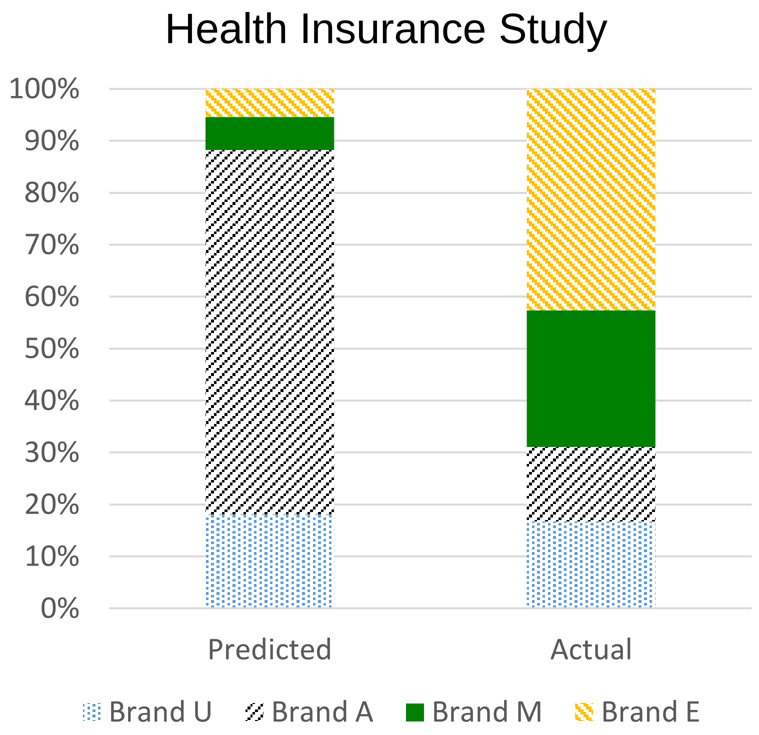
### Health Insurance Study

Many health insurance companies offer Medicare Advantage plans which are sold directly to eligible consumers (in contrast to group plans which are marketed to employers and other organizations who provide or offer insurance for employees or members). The insurance companies attempt to design benefits and premium structures that will attract profitable customers. This study tested different plan designs for Medicare Advantage products. The underlying assumption is that consumers will prefer those plans with the richest benefits relative to their needs. Thus, someone with a chronic illness that results in occasional hospitalization will look for a plan with richer hospitalization benefits, while a relatively healthy individual may be drawn to ancillary benefits such as fitness club membership. Because of the regional nature of health insurance, local Blue Cross Blue Shield carriers often have the largest market share. Possible reasons for the observed differences in market shares include awareness, consumer inertia, and time to market maturity (some of the brands are new to the local market).

The following charts display the simulated and actual market shares for each study, before any external adjustments.

[10] If using a first-choice method for simulated shares, the individual level adjustment would be applied before determining the respondent's first choice.


Patio Door Study
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Predicted
Actual
Wood
Composite
Vinyl


Health Insurance Study
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Predicted
Actual
Brand U
Brand A
Brand M
Brand E

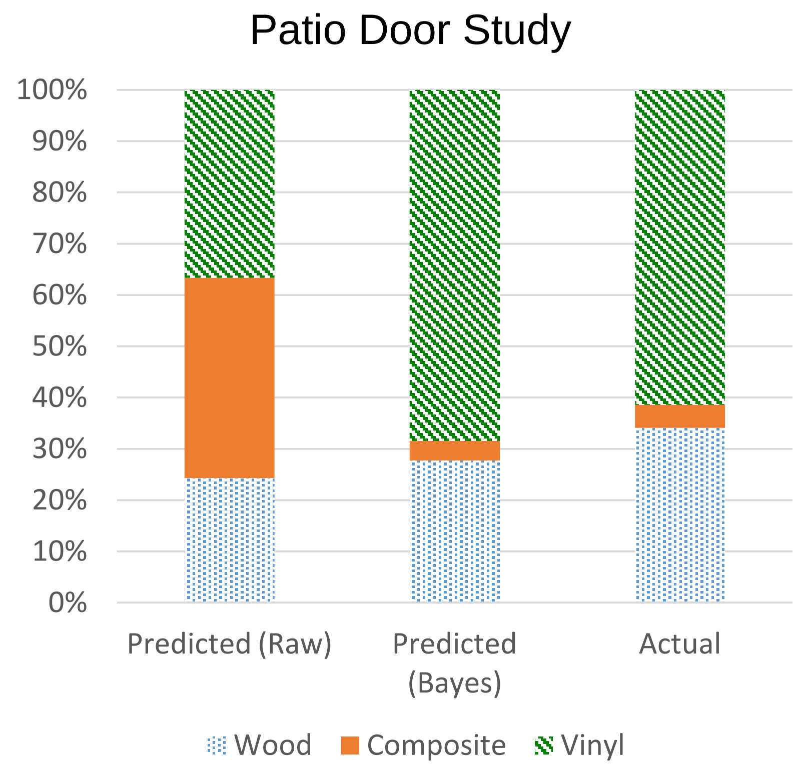
## Bayesian Adjustment of Simulated Shares

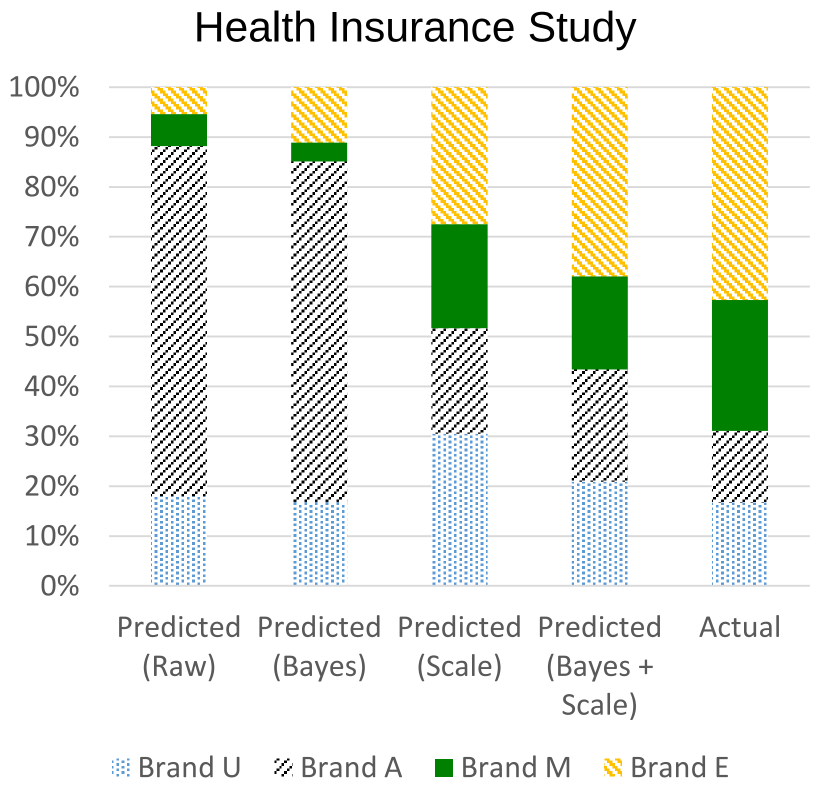
In both studies an aggregate Bayesian adjustment was applied to simulated market shares. This table illustrates the calculations for a patio door simulation. The simulation included three door options (wood, composite and vinyl) plus a “none” option.

| | Option 1 | Option 2 | Option 3 | “None” |
|---|---|---|---|---|
| Prior belief (x) | 61% | 5% | 34% | NA |
| Simulated Preference share (y) | 28.6% | 24.4% | 26.4% | 20.6% |
| Z | 25% | 25% | 25% | NA |
| Bayes calculation | 0.644 | 0.044 | 0.353 | NA |
| Rescaled* | 49.1% | 3.3% | 26.9% | 20.6% |

*Rescaling accounts for not including the “none” option in the Bayes’ calculation as there is no prior belief regarding the “none” option.

The following charts compare the unadjusted, Bayesian adjustment, and actual market shares for both studies.





For the door study, the simple Bayesian adjustment brings the simulated market shares much closer to the actual shares. In the health insurance study, one company offers a utility-dominating plan ($0 premium with relatively rich benefits). In simulation, this plan achieves about five times the observed in-market share. Because the value of **y** in the Bayesian calculation is relatively high, the Bayesian adjustment alone shrinks the share only a little. In a frictionless market, this might be an accurate representation of what will happen under conditions of 100% awareness. However, given high consumer inertia (the retention rate for the market leader, Brand E, is about 90%), it's possible that there is more "noise" in the conjoint data than in the real world. In such cases, tuning the scale factor (as applied in Sawtooth Software simulators) reduces this difference. Combining the scale factor and Bayesian adjustments brings simulated shares even closer to actual shares.

## How Does the Bayesian Adjustment Compare to Other External Adjustment Methods?

The Bayesian adjustment appears to reduce the gap between predicted and actual shares, but we might ask if it is a legitimate way to adjust simulated shares. To answer that question, we compare the Bayesian adjustment to two other adjustment methods using the patio door study. We look at three different indicators:

- Ratios of adjusted predicted shares to actual or "expected" shares (based on prior belief),
- Consistency of predictions across methods, and
- Ability to capture competitive interactions (substitution effects) between products.

Three adjustment methods are compared:

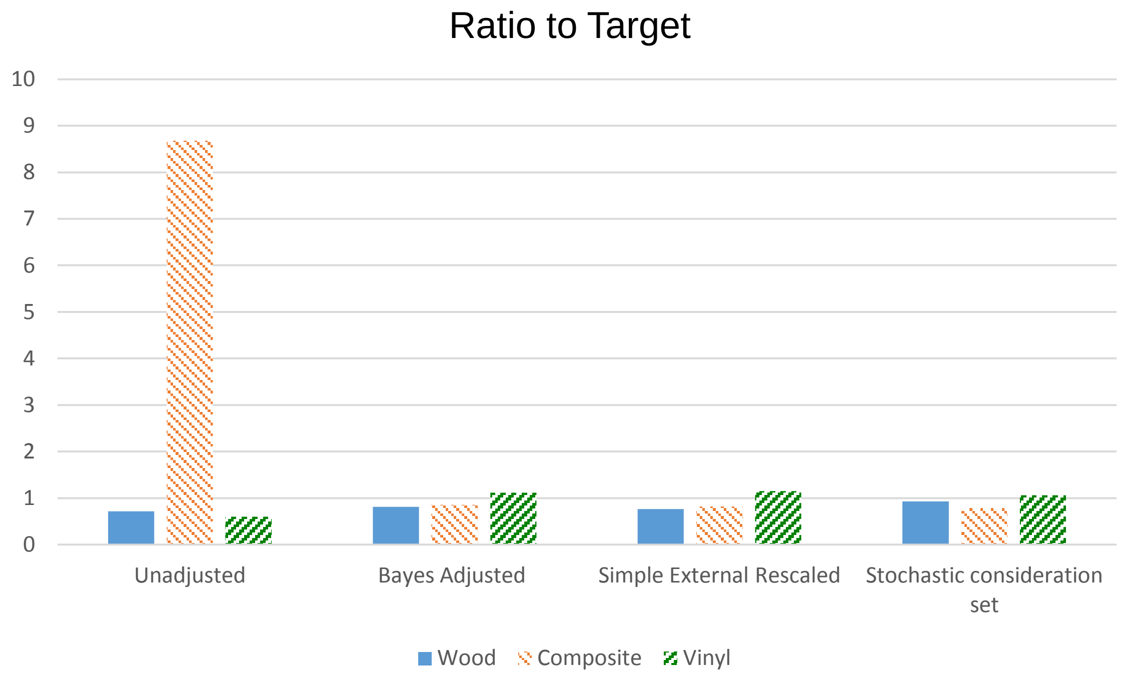
- Bayesian external adjustment (a prior belief—actual market share—is integrated with the simulated share—to arrive at a posterior prediction)
- Simple aggregate external adjustment (differential external factors such as awareness and distribution are translated into weights applied to the simulated shares)
- Stochastic consideration set formation (differential external factors such as awareness and distribution are used in a probabilistic simulation of consideration set formation).

The following table shows the calculations for the simple aggregate external adjustment.

| | **Option 1** | **Option 2** | **Option 3** | **"None"** |
|---|---|---|---|---|
| "Distribution" | 61% | 5% | 34% | |
| Simulated Preference share | 28.6% | 24.4% | 26.4% | 20.6% |
| Distribution X pref share | 17.6% | 1.1% | 9.0% | 20.6% |
| Rescaled | 50.4% | 3.1% | 25.8% | 20.6% |

The stochastic simulation method models consideration set formation as a random process dependent on the probability of being aware of the product and being able to find it when shopping (distribution). For purposes of the simulation, we assumed that the probability that a particular door type will be included in a consumer's consideration set is equal to that type's unit share of volume.[11] Using a Monte Carlo process, we generate a random number between 0 and 1 from a uniform distribution for each brand for each respondent. If that random value falls within a specified range determined by the unit share of volume, the material type is included in the individual's consideration set. This results in roughly 61% of respondents considering a vinyl door in any one simulation. To keep things simple, we multiplied the total exponentiated utility for each option by either 1 or 0, depending on whether or not it was in the consideration set, before calculating the preference shares. We then repeated this process 1,000 times and averaged across the iterations.

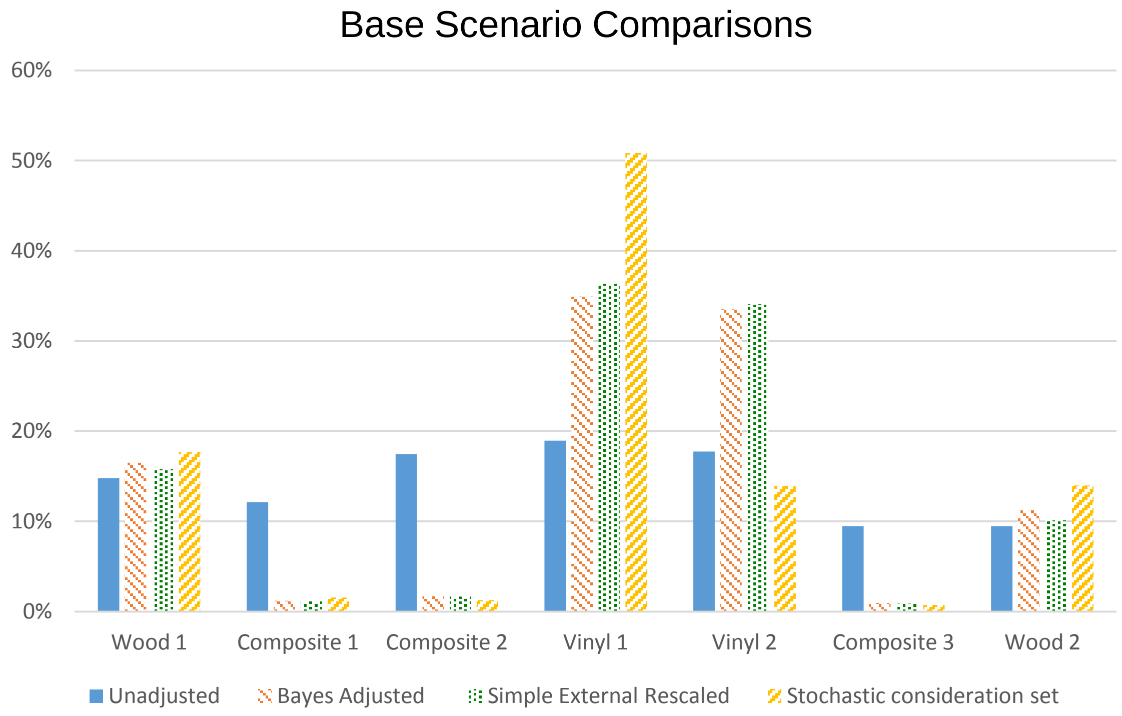
This chart compares the ratio of simulated to actual shares for each of the methods tested. We can see that for all of the adjustment methods, the ratio of simulated to actual shares is close to 1:1 for the three door types.



The following chart illustrates the degree of consistency in adjusted shares across the different methods. In this simulation, the product line-up has been expanded to capture substitution effects. For each material type, the simulation offers one "basic" product and one product with enhanced features (at a higher price). Both the Bayesian and simple aggregate adjustments yield similar adjusted shares across the different alternatives in the simulation. The stochastic simulation seems to run into some difficulty in predicting the shares for the two vinyl options; the ratio of the simulated shares for Vinyl 1 to Vinyl 2 is much different for the

[11] This is a simplifying assumption that is useful for demonstrating the differences between the methods but that would likely be inappropriate for actual simulation of consideration set formation. Among other things, we would want to separate distribution and awareness in determining whether or not a material type gets into the consideration set.

stochastic method than we see in the unadjusted shares. This particular implementation of stochastic consideration set formation appears to fail the transparency criterion.



Finally, using the same product lineup, the price for Composite Door 2 was decreased by $200 while the price for Composite Door 1 remained at $1,950. The doors are similar except for small feature differences, so we would expect Composite 2 to compete most directly with Composite 1. We might also expect to see some competition with Vinyl 1, which has a similar feature set. As the price gap between Vinyl 1 and Composite 2 shrinks, Composite 2 may become more appealing to some of those who preferred Vinyl 1.

| | Change in predicted preference share vs. base case | | | |
|---|---|---|---|---|
| | Unadjusted | Bayes Adjusted | Simple External Rescaled | Stochastic consideration set |
| Wood 1 | 0.00% | 0.11% | 0.17% | 1.90% |
| Composite 1 | -0.30% | -0.02% | -0.02% | -0.02% |
| Composite 2 | 1.48% | 0.15% | 0.16% | -0.75% |
| Vinyl 1 | -0.59% | -0.48% | -0.75% | 15.59% |
| Vinyl 2 | 0.00% | 0.22% | 0.37% | -19.63% |
| Composite 3 | -0.59% | -0.05% | -0.05% | -0.38% |
| Wood 2 | 0.00% | 0.07% | 0.11% | 3.29% |

The Bayesian and simple external adjustment methods both perform reasonably well. The stochastic simulation does not perform well. There are inexplicable swings between the two vinyl options, and Composite Door 2 actually loses some share despite the lower price.

However, there are some anomalies suggesting that the Bayesian adjustment method requires further testing across some different contexts. For example, both the Bayesian and simple external adjustments have Vinyl 2 increasing in share. The unadjusted shares do the best job of capturing the expected substitution effects.

## Recommendations and Conclusions

It is tempting to suggest that practitioners avoid making external adjustments to simulated shares whenever possible. However, there are real consequences for decision-making when simulated shares vary from actual market conditions by as much as they do in the two case studies.

The Bayesian adjustment method introduced in this paper is intuitively appealing for at least a few reasons. First, at least for those of us with a Bayesian orientation, the method offers a way to incorporate all of the information we have available into our simulation models. Second, this method is, for want of a better descriptor, self-calibrating. By that I mean that as the simulated share value (or any other input, such as our assumptions about **z** or **y**) changes, the adjustment changes in proportion to those changes. As the prior beliefs and the simulated shares converge, the adjustments decrease in magnitude. Third, if individual-level priors are available, this method might well outperform other methods.

More testing, across different contexts, is needed to confirm the promise of this method. For one thing, we need to understand the apparent anomalies in substitution effects. While the Bayesian method does a little better than the simple aggregate external adjustment in this regard, there are still concerns. Testing the method across a variety of choice models is imperative.

In the meantime, if there is a strong empirical prior like current market shares, and parameter values for factors like awareness and distribution are unknown, you may want to try this method of external adjustment.

David Bakken

## REFERENCES

Allenby, G. *et al.* (2005), "Adjusting Choice Models to Better Predict Market Behavior," *Marketing Letters*, 16:3/4, 197–208.

Bakken, D. (2006), "Agent-based Simulation for Improved Decision Making," 2006 Sawtooth Software Conference Proceedings, 83–95.

Bayes, T. (1763). Towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society, 53, 370–318. Reprinted in Biometrika, (1958), 45, 293–315.

Orme, B, and Johnson, R. (2006), "External Effect Adjustments in Conjoint Analysis," 2006 Sawtooth Software Conference Proceedings, 183–209.

Piatelli-Palmarini, M. (1994). *Inevitable Illusions*. New York: John Wiley & Sons, Inc.

Silver, N. (2012). The Signal and the Noise: Why So Many Predictions Fail—But Some Don't. New York: The Penguin Press.

# MOBILE MAXDIFF: WHAT ARE THE OPTIMAL NUMBER OF ATTRIBUTES, SCREENS AND LEVEL OF INFORMATION COMPLEXITY?

**MICHAEL PATTERSON**
*RADIUS GLOBAL MARKET RESEARCH*
**MICHAEL SMITH**
*MFOUR*

## ABSTRACT

Research presented at the 2015 Sawtooth Software Conference demonstrated that conducting MaxDiff exercises on traditional PCs (desktop/laptop), tablets and smartphones did not produce any substantive differences by device type. However, this research did not investigate various aspects associated with the presentation of content when conducting a MaxDiff study on a mobile platform.

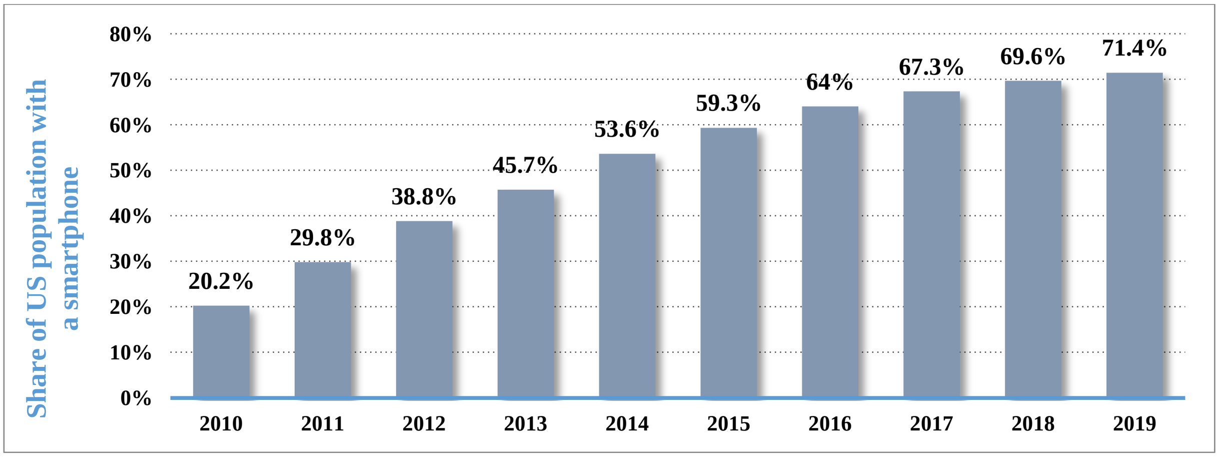
In our research, we investigated the number of items and screens to show, as well as the length of the item descriptions to determine how these different factors impact MaxDiff results when conducted on mobile devices.

We found that showing fewer items within a mobile MaxDiff study likely produces more accurate results than showing a larger number of items. Further, we did not find any impact on model accuracy as the number of screens increases. Thus, our research found that when MaxDiff studies are conducted on mobile devices, it is better to show fewer items in each question and to show more questions than presenting a greater number of items in fewer questions. We also find that minimizing the "complexity" of the information is advised.

## INTRODUCTION

Mobile devices, particularly smartphones, have become ubiquitous across the United States. It is estimated that by 2019, nearly 3 out of 4 US adults will own and use a smartphone (see Figure 1).

**Figure 1. US Smartphone Adoption Rates**



**Source:** eMarketer; US Census Bureau

Further given the prevalence of smartphones (and tablets), they are frequently used to complete online surveys (we estimate 20% to 30% are completed via mobile devices).

However, there are several notable challenges with completing surveys using mobile devices (particularly smartphones), including:

- **Smaller screen sizes**—which limit the amount and level of complexity of information that can be displayed on the screen. This can be particularly true with studies such as MaxDiff which display multiple items within each question.
- **Lower attention spans**—we would argue that individuals are often multi-tasking when using mobile devices, therefore they are often less engaged in processing information and are more easily distracted when taking surveys on a mobile device. This suggests that the presentation of information should be compelling, succinct and easy to understand/process.

Despite these challenges, mobile devices have been found to be an effective platform for completing surveys, including relatively complex questions such as MaxDiff exercises.

For example, at the 2015 Sawtooth Software conference, Jing Yeh and Louise Hanlon demonstrated that conducting MaxDiff exercises on traditional PCs such as desktops and/or laptops, tablets and smartphones did not produce any substantive differences by device type. However, they did find that smartphone surveys take longer to complete and have greater drop-off levels than do MaxDiff questions administered on PCs and tablets.

While Yeh and Hanlon (2015) showed that researchers can be confident in MaxDiff results obtained on mobile devices, they did not specifically investigate various parameters that should be kept in mind (e.g., number of items shown) when designing questions. However, other research has been conducted that can be used to provide guidance.

In one study, Chrzan & Patterson (2006) investigated the impact of displaying differing numbers of items in MaxDiff experiments that were conducted on desktops and laptops. Across three studies, they found that predictive accuracy is enhanced when researchers show 4 or 5 items in each MaxDiff question. In conclusion, they argued that it is better to utilize a larger number of questions with *fewer items* than fewer questions with more items.

In an earlier study, Orme (2005) also explored the trade-off between the number of items versus number of screens via analysis of simulated data. In his analysis, Orme found that MaxDiff experiments that presented 5 items per set provide a substantial increase in predictive accuracy versus those showing 3 items in each set. He further found that moving from 5 to 7 items per set provided little incremental impact.

Based on these previous studies, it could be surmised that displaying 4 to 5 items in each MaxDiff question might be optimal; however, given their smaller screens, it is possible that showing fewer items (3 to 4) might be optimal. But, if a researcher reduces the number of items shown in each question, there needs to be a corresponding increase in the number of tasks/screens shown in order to collect a sufficient amount of information to obtain accurate utility estimates. Increasing the number of screens could be problematic, however, given the higher drop-off levels associated with smartphones as seen in Yeh and Hanlon's (2015) study.

In previous research, one dimension that has not been investigated is the amount of descriptive content associated with the information being presented. That is, is there an effect of displaying more versus fewer words when trying to convey information to the respondents. With respect to mobile devices, it would seem that relatively short descriptions (defined as 2 to 6 words) could be easier to process on mobile devices relative to descriptive content that is more verbose (e.g., ~10+ words).

Thus, in the current study, we were interested in exploring MaxDiff on mobile devices to determine whether there is an “optimal”:

- Number of items to show in each MaxDiff question
- Number of screens to display
- Amount of descriptive content to present to respondents

## Research Design

Respondents were recruited via MFour’s mobile panel which consists of individuals who complete surveys via an app on Apple and Android mobile devices (smartphones & tablets). An app does not exist for PCs; therefore we are certain that respondents did not complete surveys via desktops or notebook computers. The survey took respondents approximately 15 minutes to complete. To qualify, respondents were required to be decision makers for technology purchases within their homes. During the MaxDiff exercise, respondents were asked to evaluate the importance of various features and capabilities when purchasing a PC (either a desktop or notebook) or tablet computer for use in their home.

Prior to the MaxDiff exercise, respondents were randomly assigned to one of 8 experimental conditions as shown in Table 1.

**Table 1: Experimental Conditions**

| Experimental Condition | # of items shown in each set | # of screens shown | Amount of descriptive Information |
|---|---|---|---|
| 1 | 3 | 8 | succinct |
| 2 | 5 | 8 | succinct |
| 3 | 7 | 8 | succinct |
| 4 | 4 | 6 | succinct |
| 5 | 4 | 9 | succinct |
| 6 | 4 | 12 | succinct |
| 7 | 4 | 8 | succinct |
| 8 | 4 | 8 | longer |

The first three experimental groups allow us to investigate the impact, if any, of varying the number of items shown in each MaxDiff question. In this case, respondents either saw 3, 5, or 7 items in the MaxDiff question while the number of questions asked was held constant at 8 for each respondent. Respondents in these three groups were also shown content with descriptions that consisted of 2 to 6 words (defined as "succinct" descriptors).

Experimental conditions 4 through 6 allowed us to look at the influence of the number of MaxDiff questions answered by respondents. Respondents were shown 6, 9, or 12 MaxDiff questions in these groups. Four items were consistently shown in each of the questions and all of their content was composed of "short" descriptions.

The final two groups varied the amount of information used to describe the product features. In the "short" condition, items consisted of 2 to 6 words, while those in the "long" condition were show descriptions that consisted of 7 to 20+ words. The items tested in the two groups are shown in Table 2.

**Table 2. Items Tested in the MaxDiff Exercise**

| **Succinct Descriptions** |
|---|
| Device price |
| Battery life or power consumption |
| Overall visual experience |
| CPU processor brand |
| Size/weight/form factor |
| Device manufacturer |
| Warranty/after sale service and support |
| Storage Capacity |
| Number of ports or external connections |
| **Longer Descriptions** |
| Device price (price of the device excluding tax) |
| Battery life or power consumption (how long the device will run on battery, or amount of power it consumes) |
| Overall visual experience (e.g., resolution and graphic clarity) |
| CPU processor brand (the manufacturer of the processor) |
| Size/weight/form factor (the physical characteristics of the device) |
| Device manufacturer (the brand of the device) |
| Warranty/after sale service and support (the extent to which service and support are available after the purchase) |
| Storage Capacity (how much internet storage is available on the device—either a hard drive or SSD) |
| Number of ports or external connections (includes ports such as USB, HDMI, etc.) |

A total of 2,002 completes were obtained in the US. The completed surveys were evenly distributed across the 8 experimental conditions so that each group had 250 completes with the exception of Experimental Condition #2 which had 252 completed surveys.

## PLANNED COMPARISONS

Several planned comparisons were conducted:

### Drop-Off Rates

We measured the proportion of respondents who suspended the survey by failing to complete the MaxDiff exercise once they began. This analysis involved conducting a two-proportion z-test to test each of the experimental conditions against one another to determine if they were significantly different.

### Task Length

We assessed the average amount of time (in seconds) it took individuals to complete the MaxDiff exercise. We would expect as the number of screens increase, the task length should also increase. It is also possible as the number of items increase, the time will also increase. Analysis of Variance (ANOVAs) with post-hoc tests were performed to compare the means within each group of experimental conditions (e.g., conditions 1–3, 4–6, 7 and 8).

### McFadden's Pseudo $R^2$

McFadden's pseudo $R^2$ (also known as Percent Certainty) was computed to judge the internal fit within each respondent in each of the experimental condition groups. In this case, McFadden's pseudo $R^2$ was calculated as 1 - (LN RLH / LN (1/C)) where RLH is the root likelihood and C is the number of items shown in each MaxDiff exercise. The pseudo $R^2$ were then analyzed via ANOVAs with post-hoc tests to determine if there were differences between experimental conditions.

### Predictive Validity

To assess predictive validity, respondents were randomly shown two of six MaxDiff holdout questions, each of which contained four items. The holdout questions were always the last two questions asked in the MaxDiff series. Those in the "longer" description group saw long descriptive items; all others saw succinct items. Two measures of predictive validity were examined.

- **Hit Rate:** The percentage of time we could predict each individual respondent's best and worst responses based on the MaxDiff utilities. Higher hit rates are preferred over lower hit rates.
- **Mean Absolute Deviation (MAD):** Allows us to measure how well the aggregate best and worst share predictions match actual shares. Smaller MAD is better than higher.

ANOVAs with post-hoc tests were performed to compare hit rates and MAD within groups of experimental conditions.

### Parameter Equivalence

We examined the utilities from each of the conditions to determine if there were significant differences. The utilities were transformed via zero-centered internal scaling to remove differences in the scale factor that might exist across groups. Multivariate ANOVAs

(MANOVAs) with post-hoc tests were computed to compare the utility vectors across experimental condition groups.

### Task Perceptions

To assess respondents' perceptions of the survey experience, we asked three questions at the conclusion of the survey to determine respondents' *satisfaction* with the survey, the *ease* of the survey to complete and how *accurate* they felt their answers were.

## RESULTS

### Drop-off Rates & Task Length

As the number of screens increases, the drop-off rate also increases. Significantly more respondents dropped off when shown 9 or 12 screens vs. only 6 screens. As the number of items increases from 3 to 5, there is also an increase in the dropout rate, although the difference is not significantly different.

For task length, as expected, as the number of items and screens increase, the time to complete increases. The same is true for longer text vs. succinct text.

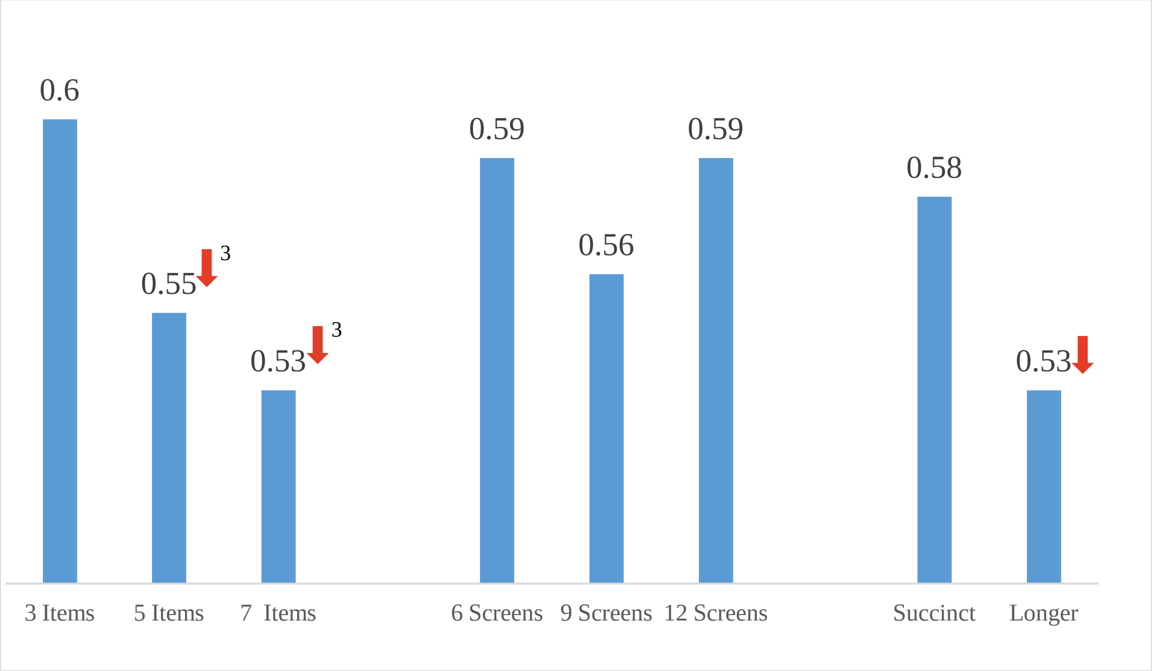
**Table 3: Drop-off Rates & Task Length**

| Experimental Condition | Drop-off Rate | Task Length (s) |
|---|---|---|
| 3 Items | 7% | 209↓[7] |
| 5 Items | 12% | 235 |
| 7 Items | 12% | 241 |
| 6 Screen | 6%↓[9,12] | 183 |
| 9 Screen | 11% | 228 |
| 12 Screen | 13% | 258 |
| Succinct | 9% | 201↓ |
| Longer | 10% | 257 |

### McFadden's pseudo $R^2$

When examining McFadden's pseudo $R^2$, higher values suggest that we have more accurate utility estimates and hence are better able to predict respondent's preferences. The results demonstrate that showing 3 items leads to a significantly higher pseudo $R^2$ in comparison to displaying 5 or 7 items. Providing a shorter item description also produces a higher pseudo $R^2$ relative to longer descriptions. This suggests that showing fewer items that are more succinct in nature will lead to more accurate utility estimates that provide better fitting MaxDiff models.

**Figure 2. Pseudo R$^2$**



## Hit Rates & Mean Absolute Differences (MAD)

We find that displaying 7 items leads to a significantly lower hit rate in comparison to 3 or 5 items. Further, displaying longer descriptions results in a lower hit rate in comparison to more succinct items. There is no difference in hit rates based on the number of screens shown.

There are no differences across the experimental conditions in terms of the mean absolute differences.

**Table 5. Hit Rate and Mean Absolute Deviation (MAD)**

| Experimental Condition | Hit Rate | MAD |
|---|---|---|
| 3 Items | 70% | 4% |
| 5 Items | 70% | 5% |
| 7 Items | 61% ↓ $^{3,5}$ | 5% |
| 6 Screen | 68% | 5% |
| 9 Screen | 71% | 4% |
| 12 Screen | 72% | 4% |
| Succinct | 57% | 5% |
| Longer | 54% ↓ | 5% |

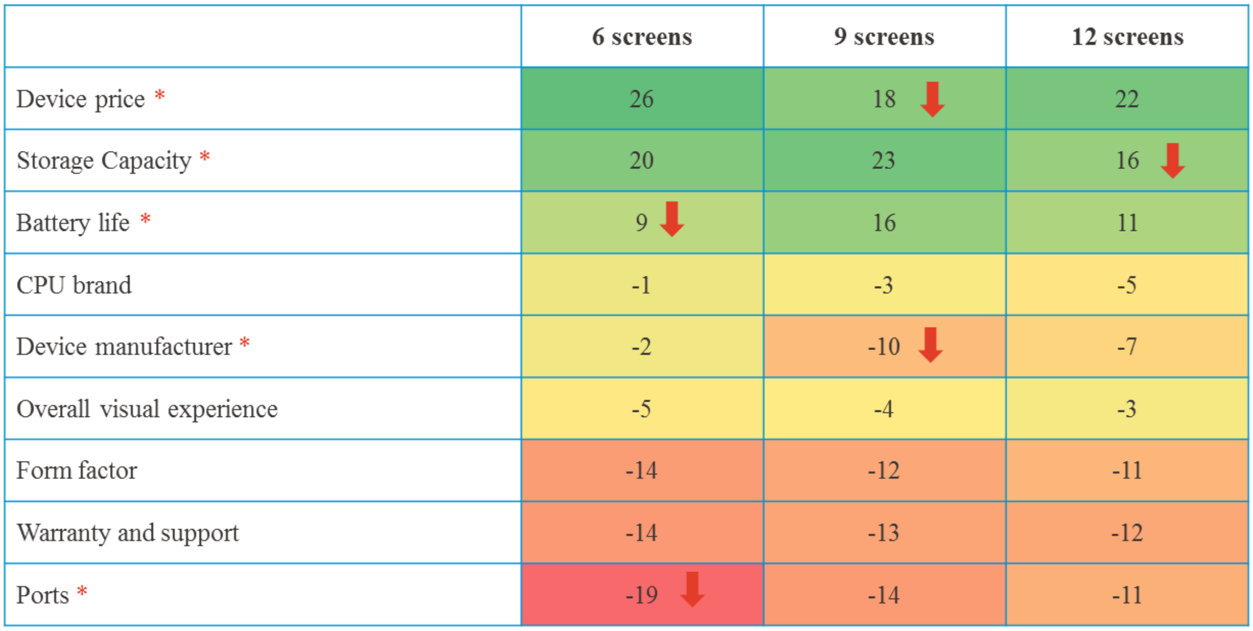
Testing for differences in parameters reveals that there are no statistically significant differences among the three number of items conditions.

**Table 6. Utilities by Number of Items Shown**

| | 3 items | 5 items | 7 items |
|---|---|---|---|
| Device price | 24 | 18 | 16 |
| Storage Capacity | 19 | 18 | 19 |
| Battery life | 10 | 11 | 10 |
| Overall visual experience | 2 | 1 | 0 |
| CPU brand | -3 | -1 | -1 |
| Device manufacturer | -11 | -5 | -6 |
| Warranty and support | -13 | -10 | -14 |
| Form factor | -13 | -15 | -11 |
| Ports | -14 | -16 | -13 |

Looking at the screen conditions, we do find significant differences in the parameter estimates. However, from a managerial standpoint, we would draw similar conclusions concerning the directional preferences of the items regardless of group. That is, looking at the rank ordering of the utilities, we see that the appeal of the items is nearly identical across the three conditions.

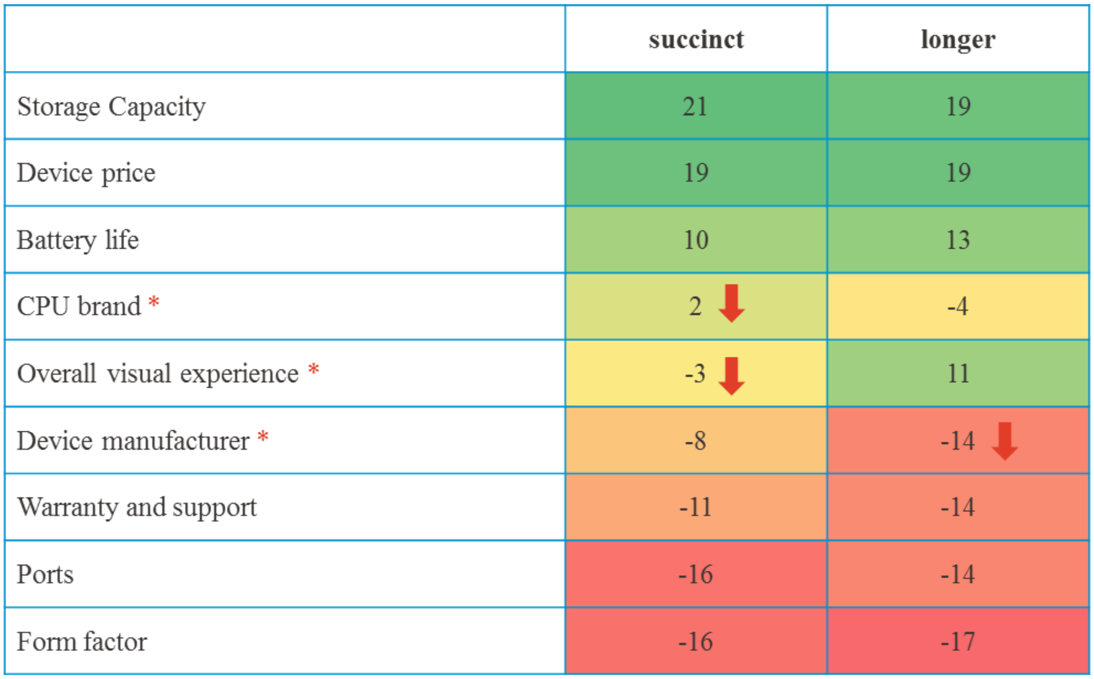
**Table 7. Utilities by Number of Screens Shown**



| | 6 screens | 9 screens | 12 screens |
|---|---|---|---|
| Device price * | 26 | 18 ⬇ | 22 |
| Storage Capacity * | 20 | 23 | 16 ⬇ |
| Battery life * | 9 ⬇ | 16 | 11 |
| CPU brand | -1 | -3 | -5 |
| Device manufacturer * | -2 | -10 ⬇ | -7 |
| Overall visual experience | -5 | -4 | -3 |
| Form factor | -14 | -12 | -11 |
| Warranty and support | -14 | -13 | -12 |
| Ports * | -19 ⬇ | -14 | -11 |

We also find significant difference in our parameter estimates based on the amount of content displayed. While most of the differences are relatively modest, the one that is most notable is related to “Overall visual experience” where we find higher utilities for those in the longer description condition. In this case, we could surmise that the description provided additional information leading respondents to judge that this factor was more important.

**Table 8. Utilities by Amount of Information**



| | succinct | longer |
|---|---|---|
| Storage Capacity | 21 | 19 |
| Device price | 19 | 19 |
| Battery life | 10 | 13 |
| CPU brand * | 2 ⬇ | -4 |
| Overall visual experience * | -3 ⬇ | 11 |
| Device manufacturer * | -8 | -14 ⬇ |
| Warranty and support | -11 | -14 |
| Ports | -16 | -14 |
| Form factor | -16 | -17 |

### Task Perceptions

There are no significant differences in terms of satisfaction levels, perceived ease of completion or accuracy based on the number of items seen nor the number of screens shown. The only significant difference is among those in the longer item description condition who express higher satisfaction levels than the succinct group, however, the effect is relatively modest.

**Table 9. Satisfaction, Ease of Completion & Accuracy**

| Experimental Condition | Satisfaction | Ease of Completion | Accuracy |
|---|---|---|---|
| 3 Items | 1.89 | 1.90 | 1.80 |
| 5 Items | 1.85 | 1.92 | 1.80 |
| 7 Items | 1.95 | 2.00 | 1.83 |
| 6 Screen | 1.88 | 1.91 | 1.75 |
| 9 Screen | 1.90 | 1.94 | 1.79 |
| 12 Screen | 1.90 | 1.98 | 1.82 |
| Succinct | 1.87↓ | 1.83 | 1.78 |
| Longer | 1.98 | 2.06 | 1.88 |

## Discussion

In this research, we find that showing fewer items (ideally 3, but up to 5) within a MaxDiff study conducted on mobile devices is likely to produce more accurate results (when judged by holdout hit-rates and pseudo $R^2$ values) versus studies that show a larger number of items (7 and perhaps more). Further, there seems to be no impact on model accuracy as the number of screens increases (at least up to 12 screens).

Thus, like Chrzan & Patterson (2006), we would argue that with MaxDiff studies conducted on mobile devices it is advised to show fewer items in each question and to show more questions than presenting respondents with a greater number of items in fewer questions.

Further, we also find that as much as possible, the verbiage associated with the items presented to respondents should be minimized thus resulting in more accurate results (again as judged by hit-rates and pseudo $R^2$). Obviously, items need to be adequately explained so that respondents understand them sufficiently which suggests that pre-testing of items is warranted to ensure proper understanding.

Michael Patterson

Michael Smith

## References

Chrzan, K., Patterson, M. (2006). Testing for the Optimal Number of Attributes in MaxDiff Questions. *Proceedings of the 2006 Sawtooth Software Conference,* Delray Beach, Florida. March 2006.

Orme, B. (2005) Accuracy of HB Estimation in MaxDiff Experiments. *Sawtooth Software Research Paper*, http://www.sawtoothsoftware.com/download/techpap/maxdacc.pdf

Yeh, J., Hanlon, L. (2015). MaxDiff on Mobile. *Proceedings of the 2015 Sawtooth Software Conference,* Park City, UT. March 2015.

# CHOICE-BASED CONJOINT IN A MOBILE WORLD— HOW FAR CAN WE GO?

***CHRIS MOORE***
***CHRISTIAN NEUERBURG***
*GFK*

## ABSTRACT

With more than two-thirds of panel respondents having now used a tablet or smartphone to answer surveys it is just as important as ever to understand what effect this has on the results of conjoint studies. Past research into the effect of conducting conjoint surveys on a mobile device has typically concentrated on how to simplify the conjoint task and have tested a limited number of designs. To expand knowledge in this research area we conducted the most comprehensive study known using an 18 split-sample design and more than 6,800 respondents to evaluate what effect different conjoint designs have on conjoint data. Rather than asking the question of how simple do we need to make design, the question is how complex a design can we show and still obtain robust results? Through this research we are able to conclude that respondents can comfortably cope with complex designs with very little, if any, degradation on the robustness of the data. Interesting findings were also captured regarding the environment respondents take mobile surveys in, which conflict with popular perception and the demographic composition of these respondents. Analysis was also conducted to look at the effect of showing concepts vertically rather than the more traditional horizontal layout, which showed only minor differences.

## BACKGROUND

Over recent years the usage of mobile devices (defined in this paper as a smartphone or tablet) makes up a significant minority of the interviews collected on market research access panels. Internal research on past projects conducted in 2015 concluded that on any one survey the proportion of panelists answering a survey on a mobile device was:

**Proportion of Respondents Using Mobile Devices**

| | |
|---|---|
| Germany | 12% |
| UK | 14% |
| USA | 21% |

Source: GfK

This shift towards using mobile devices to answer surveys also has a significant impact on dropout rates. In a comparison of dropout rates for those respondents that answer on a desktop/laptop versus a mobile device, those answering on a smartphone are more likely to drop out of the survey. Figures of dropout rates are shown below.

**Dropout Rates**

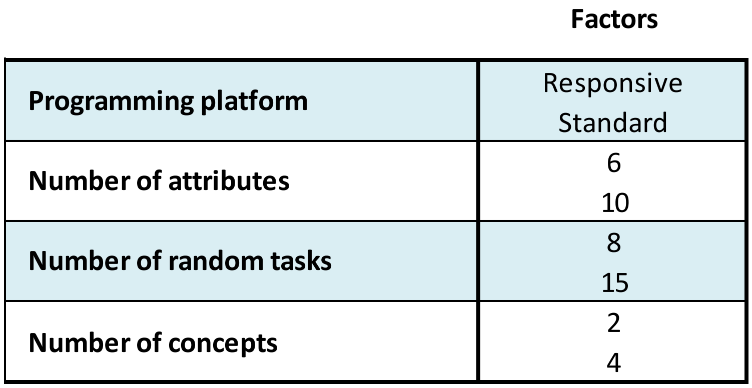
| | Desktop/Laptop | Tablet | Smartphone |
|---|---|---|---|
| Germany | 6% | 20% | 22% |
| UK | 6% | 14% | 22% |
| USA | 14% | 31% | 30% |

In the research conducted for this paper, of the 6,866 respondents that were interviewed, over 91% had previously used a smartphone or tablet to answer a market research survey (54% had previously used a smartphone and 61% had previously used a tablet). Due to an increasing use of mobile devices, most market research agencies now implement responsive web designs to provide a better user experience for respondents answering a survey on a mobile device. However, these responsive designs to date have typically been limited to standard survey questions rather than the conjoint exercise itself.

Two papers that were presented at the 2013 Sawtooth Software conference (Diener *et al.*, White) investigated whether there are any differences in conjoint results when a conjoint survey had been conducted via a mobile device versus more traditional means (laptop or PC). Both papers found little evidence of any differences but the research was restricted in terms of the number of experimental conditions that were tested, for example, the number of tasks, the number of concepts and the number of attributes contained within the design. The focus instead was more on testing simplifying strategies that may be needed by the researcher in order to obtain comparable conjoint results.

As these papers showed evidence that there is little need to simplify conjoint designs, the primary objective of this paper is to systematically research the effect of different conjoint design settings within a mobile environment and to recommend where possible the best combination of design parameters to conduct mobile CBC experiments. In order to test this, a 16 split sample design was set up where the experimental design conditions differed by cell. The experimental conditions tested were:

**Experimental Factors Tested**



| | **Factors** |
|---|---|
| **Programming platform** | Responsive<br>Standard |
| **Number of attributes** | 6<br>10 |
| **Number of random tasks** | 8<br>15 |
| **Number of concepts** | 2<br>4 |

In addition to these 16 mobile cells, an additional 2 cells were included where respondents completed the conjoint experiment on a fixed device (desktop or laptop), which allowed, as a secondary objective, comparisons to be made between device types. To be able to also make some direct comparisons between previous work in this area, the study area used for the conjoint experiment was based on the preference for tablets.

## What We (Think We) Know about "Mobile" Respondents

When we think of a respondent answering a survey on a mobile device, the image of what one thinks is happening is not necessarily what is actually happening in real-life. The immediate thought is that someone conducting a survey on a mobile device is either in a public place or in a busy and/or distracting environment, and as a consequence it is not possible to give their full attention to the stimuli that is required to give sensible and consistent answers to the conjoint tasks. They are also viewing the questions/conjoint tasks on a much smaller screen so the survey experience is compromised and these people may be very time conscious so are only willing to spend a short time to answer the entire survey. Compare this to someone who is answering on their PC/laptop and is at home in a quiet and relaxed environment. We assume that they are giving the survey their full attention, have a good user experience because of the much bigger screen size and are willing to spend more time to answer the survey.

Within this research a number of questions were asked about the environment that the survey was taken, in addition to survey experience questions. Analysis that will be presented later in the paper will show that the reality is that respondents answering on mobile devices are just as likely to be at home in a quiet a relaxed atmosphere and despite the smaller screen size there is little to suggest that their user experience has been significantly affected.

In terms of what we know about differences between respondents answering on a mobile device and those answering on a fixed device (Desktop/laptop) previous papers (Diener *et al.* 2013, White 2013) have shown that between device types:

- Conjoint utilities are highly correlated.
- Holdout accuracy is comparable.
- Mean Absolute Errors (MAE) are comparable.
- Dropout rates are comparable.
- The user experience is slightly less enjoyable for mobile users.
- Conjoint exercises taken on mobile take slightly more time to complete.

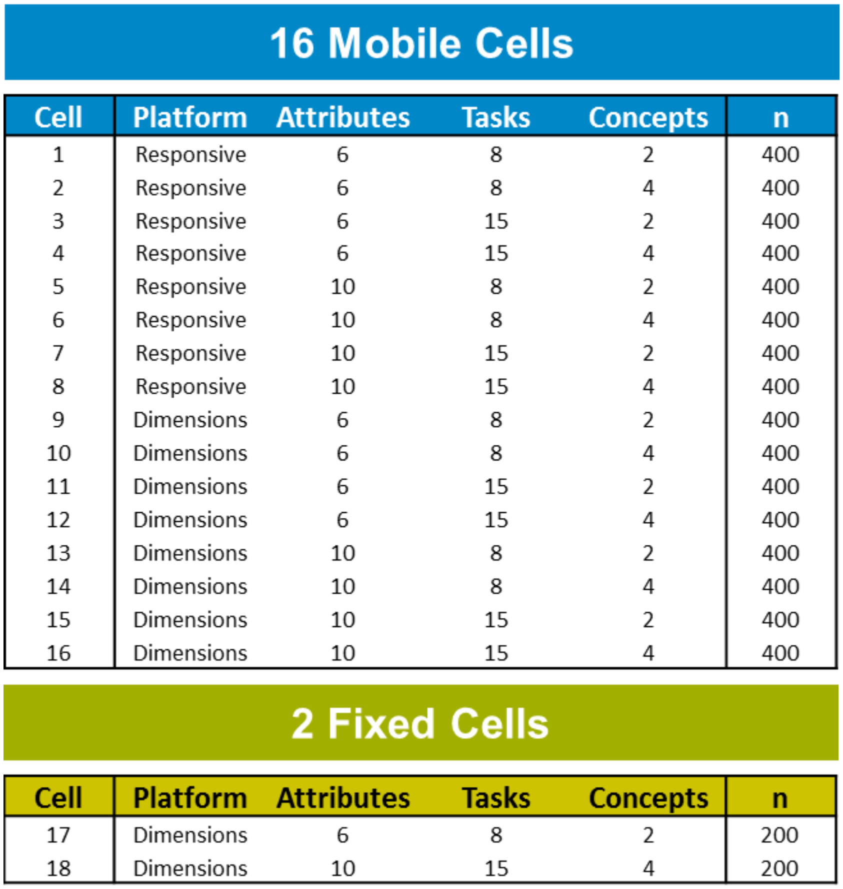
This research has shown similar findings to most of these points so it adds to the body of evidence that indicates that there is little, if any systematic bias when answering conjoint studies on a mobile device.

We also know from demographic analysis of respondents answering surveys on a mobile device that they tend to be younger and are more likely to be female.

## Study Design

In order to systematically test different design conditions, a 16 monadic cell design was set up, each testing a specific design configuration (Figure 1). A sample size of N=400 was assigned to each cell, and soft quotas were included to ensure a minimum of N=175 respondents answered on smartphone and N=175 respondents answered on a tablet (quotas were not needed as each cell had c.50% of each device type). No other quotas were included in the survey. As a benchmark, two additional cells were included where respondents answered via a desktop or laptop. The most simple and most complex combinations of experimental factors were chosen for these fixed cells and there was a sample of 200 in each of these cells.

**Figure 1. Experimental Designs Tested**



16 Mobile Cells

| Cell | Platform | Attributes | Tasks | Concepts | n |
|---|---|---|---|---|---|
| 1 | Responsive | 6 | 8 | 2 | 400 |
| 2 | Responsive | 6 | 8 | 4 | 400 |
| 3 | Responsive | 6 | 15 | 2 | 400 |
| 4 | Responsive | 6 | 15 | 4 | 400 |
| 5 | Responsive | 10 | 8 | 2 | 400 |
| 6 | Responsive | 10 | 8 | 4 | 400 |
| 7 | Responsive | 10 | 15 | 2 | 400 |
| 8 | Responsive | 10 | 15 | 4 | 400 |
| 9 | Dimensions | 6 | 8 | 2 | 400 |
| 10 | Dimensions | 6 | 8 | 4 | 400 |
| 11 | Dimensions | 6 | 15 | 2 | 400 |
| 12 | Dimensions | 6 | 15 | 4 | 400 |
| 13 | Dimensions | 10 | 8 | 2 | 400 |
| 14 | Dimensions | 10 | 8 | 4 | 400 |
| 15 | Dimensions | 10 | 15 | 2 | 400 |
| 16 | Dimensions | 10 | 15 | 4 | 400 |

2 Fixed Cells

| Cell | Platform | Attributes | Tasks | Concepts | n |
|---|---|---|---|---|---|
| 17 | Dimensions | 6 | 8 | 2 | 200 |
| 18 | Dimensions | 10 | 15 | 4 | 200 |

The experimental conditions were defined as:

### Platform:

The non-conjoint element of the main questionnaire was conducted using SPSS Dimensions© and it had been optimized for use on a mobile device. For the conjoint part of the questionnaire this differed, where in one experimental condition the conjoint task was still taken within the same Dimensions platform (we refer to this as “Dimensions”) while in the other condition a responsive method called Angular JS (we refer to this as “Responsive”) was implemented. This involved routing out of the main questionnaire to conduct the conjoint experiment before routing back in to the Dimensions platform. The Angular JS method is a framework which is used for creating responsive single page applications and consists of a collection of libraries and directives. The main components of the framework are:

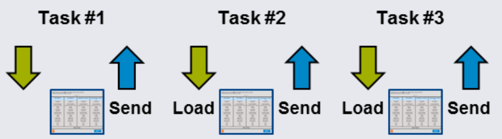
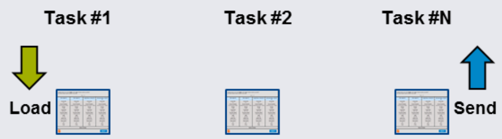
- **Angular** (by Google) is a fast and simple JavaScript framework that allows the creation of single page applications in an easy an efficient manner.

- **Twitter Bootstrap** delivers the responsiveness to the platform by automatically repositioning and resizing based on the screen resolution that the respondent is conducting the survey on.
- **Yeoman/Bower/Grunt** are tools that are combined and used as an app creator which delivers a test framework and test server.

The combination of these techniques provides a multi-device application with the same layout quality as Flash but with the advantage that it allows complex surveys to be conducted via smartphones and tablets. An advantage of this method over standard interviewing platforms like Dimensions or Confirmit is that it is a single page application process and as such the interviewing time is reduced. With a platform such as Dimensions, there involves a significant amount of server communication in order to process the result that a respondent has given for a task and then to upload the new task. With the Angular JS system, the conjoint tasks are loaded onto a single page application that is uploaded prior to commencing the conjoint part of the survey so there is no lag between submitting an answer to a task and the next task appearing.

The presentation of the concepts will also differ. With the responsive Angular JS platform, the concepts will either be alongside one another (horizontal layout) or below one another (vertical layout) depending on the orientation and size of the screen that the respondent is using. In a Dimensions platform the concepts always appear alongside to one another regardless of orientation and size of screen. Appendix A shows example screenshots of the layout of the conjoint tasks by device and orientation. Figure 2 shows a summary of the differences between Dimensions and Angular JS.

**Figure 2. Comparison of Dimensions and Angular JS**

| | Dimensions | Angular JS |
|---|---|---|
| **Responsive-ness** | Automatically adapts to different screen sizes but cannot adapt to different device orientations | Automatically adapts to different screen sizes and device orientations |
| **Concept Presentation** | Landscape: Next to each other<br>Portrait: Next to each other | Landscape: Next to each other<br>Portrait: Below each other |
| **Concept Selection** | By hitting a check box | By clicking the whole concept card |
| **Server Communi-cation** |   Task #1<br>Task #2<br>Task #3<br>Send<br>Load<br>Send<br>Load<br>Send |  |

## Attributes:

A six- and a ten-attribute design were used across the cells. The attributes and levels were chosen based on actual retail sales data and reviewing the technical descriptions of the leading 50 tablets sold in the UK in 2014. See Appendix B for a description of the attributes and levels.

### Tasks:

Designs with eight or fifteen random tasks were created. In addition to the random tasks, two additional holdout tasks were included, meaning in total, ten or seventeen tasks were shown to respondents.

### Concepts:

Designs either had two or four concepts, plus a None option for respondents to make a single choice from.

The designs were generated in Sawtooth Software SSI Web v8.3.10. For each design, 50 versions of the tasks were generated using a balanced overlap algorithm.

## Estimation Procedure

The analysis was conducted using Sawtooth Software's CBC/HB v5.5.3 and the analysis was run separately for each device type (smartphone and tablet) within cell. This resulted in 34 separate runs (16 cells x 2 device types + 2 fixed cells) and was done to ensure that any differences between respondents were not influenced by the device type. A part-worth estimation procedure was used and no covariates or constraints were included in the estimation procedure.

Due to the split sample design, in order to make comparisons across cells it is important to ensure the demographic balance of the cells. Through the random fallout of interviews, the demographic balance was very similar across the mobile cells but a RIM weighting procedure in SAS was implemented where the target weights were set as the demographic composition across the entire sample. Each cell (including the fixed cells) was weighted on:

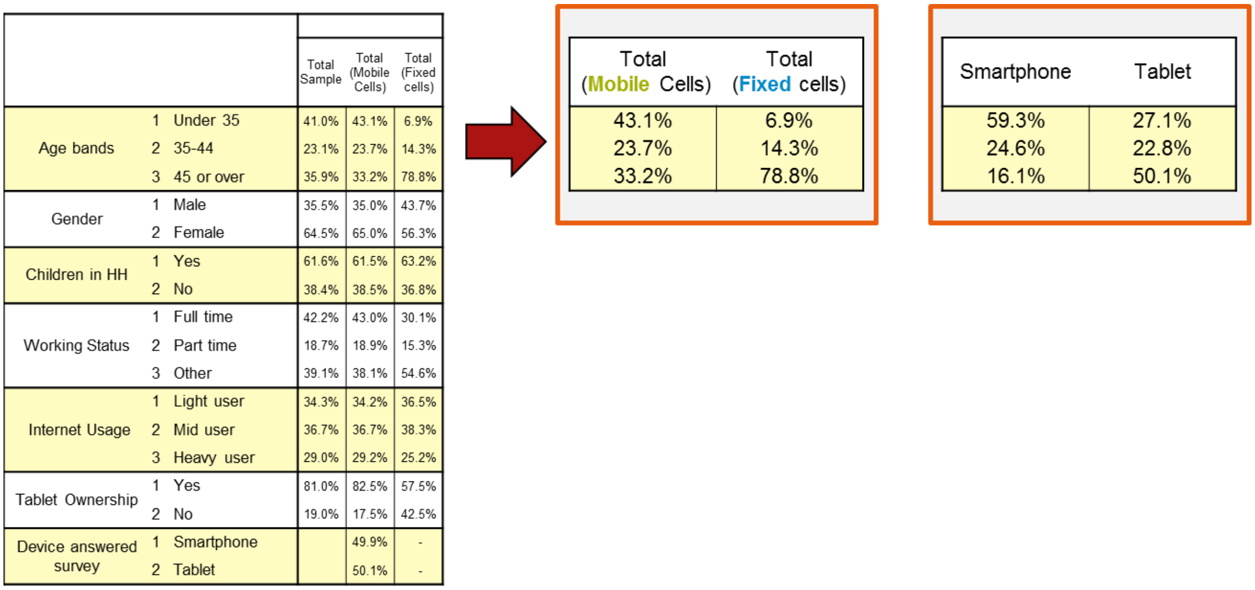
- Age
- Gender
- Children in HH
- Working status
- Internet usage
- Tablet ownership
- Device type (used to conduct survey)*
  * For the two fixed cells the device types was not applicable.

## Results

### Sample demographics

A review of the unweighted demographic data by cell showed large differences between the mobile and fixed samples in terms of age distribution and gender. This backs up previous research done in this area which has shown similar patterns. Figure 3 shows further details of the demographic differences between the mobile and fixed cells.

### Figure 3. Demographic Breakdown of Mobile and Fixed Cells



| | | | Total Sample | Total (Mobile Cells) | Total (Fixed cells) |
|---|---|---|---|---|---|
| Age bands | 1 | Under 35 | 41.0% | 43.1% | 6.9% |
| | 2 | 35-44 | 23.1% | 23.7% | 14.3% |
| | 3 | 45 or over | 35.9% | 33.2% | 78.8% |
| Gender | 1 | Male | 35.5% | 35.0% | 43.7% |
| | 2 | Female | 64.5% | 65.0% | 56.3% |
| Children in HH | 1 | Yes | 61.6% | 61.5% | 63.2% |
| | 2 | No | 38.4% | 38.5% | 36.8% |
| Working Status | 1 | Full time | 42.2% | 43.0% | 30.1% |
| | 2 | Part time | 18.7% | 18.9% | 15.3% |
| | 3 | Other | 39.1% | 38.1% | 54.6% |
| Internet Usage | 1 | Light user | 34.3% | 34.2% | 36.5% |
| | 2 | Mid user | 36.7% | 36.7% | 38.3% |
| | 3 | Heavy user | 29.0% | 29.2% | 25.2% |
| Tablet Ownership | 1 | Yes | 81.0% | 82.5% | 57.5% |
| | 2 | No | 19.0% | 17.5% | 42.5% |
| Device answered survey | 1 | Smartphone | | 49.9% | - |
| | 2 | Tablet | | 50.1% | - |

| Total (Mobile Cells) | Total (Fixed cells) |
|---|---|
| 43.1% | 6.9% |
| 23.7% | 14.3% |
| 33.2% | 78.8% |

| Smartphone | Tablet |
|---|---|
| 59.3% | 27.1% |
| 24.6% | 22.8% |
| 16.1% | 50.1% |

An interesting finding is that when the mobile sample is split between those that answered the survey on a smartphone and those that answered on a tablet, the age distribution differs significantly again. Tablet users are more aligned to the fixed cells in terms of age demographic and are much older that those that answer on a smartphone.

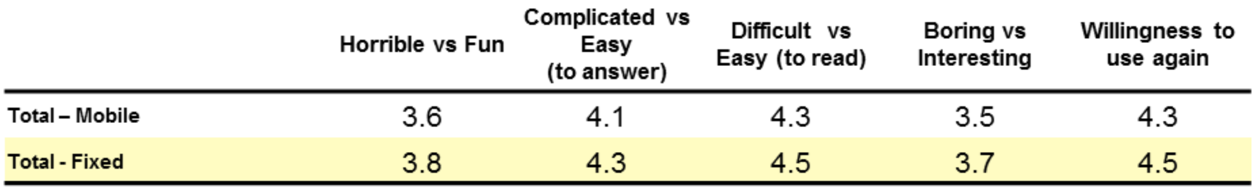
## User Experience

Four user experience questions were included in the survey after the conjoint section to understand differences in user experience across the different experimental conditions. An additional question about willingness to answer future surveys using the same device was also asked. The questions were asked on a 5-point anchored scale.

1. Horrible (1) vs. Fun (5)
2. Complicated to answer (1) vs. Easy to answer (5)
3. Difficult to read (1) vs. Easy to read (5)
4. Boring (1) vs. Interesting (5)
5. Unwilling to use device again (1) vs. Willing to use device again (5)

When comparing results for the mobile cells against the fixed cells the latter consistently scored higher by c.0.2 points on average on all statements (Figure 4). However, the results suggest there is little difficulty in completing the exercise on a mobile device with readability averaging 4.3 (out of 5) and willingness to use device type again also averaging 4.3. While the results are statistically significant between the fixed and mobile cells (due to the large sample size) the results suggest that there are no real concerns regarding conducting surveys on mobile devices. When looking at the mobile cells only, the biggest difference is between the device types (smartphone vs. tablet) where the tablet consistently scored higher by up to 0.2 points.

**Figure 4. User Experience Questions**



| | Horrible vs Fun | Complicated vs Easy (to answer) | Difficult vs Easy (to read) | Boring vs Interesting | Willingness to use again |
|---|---|---|---|---|---|
| **Total – Mobile** | 3.6 | 4.1 | 4.3 | 3.5 | 4.3 |
| **Total - Fixed** | 3.8 | 4.3 | 4.5 | 3.7 | 4.5 |

Differences across the four sets of experimental conditions were minimal (generally less than 0.1 points in mean score). For the platform type, number of attributes and number of tasks, differences were all less than 0.1 and it is only with the number of concepts where differences seen are between 0.1-0.2 in favour of the 2 concept design. A comparison was also conducted to look at the different combinations of concepts/tasks. While it would be expected that a 2 concept/8 task design would have better ratings than the 4 concept/15 task design, the total amount of information shown is similar for the 2 concept/15 task design and the 4 concept/8 task design so any differences may lead to a recommendation to use one over the other. Like the other experimental conditions, while the scores across the conditions were very similar, the 2 concept/15 task design combination had better ratings on statements 2 and 3 and comparable ratings for the other statements.

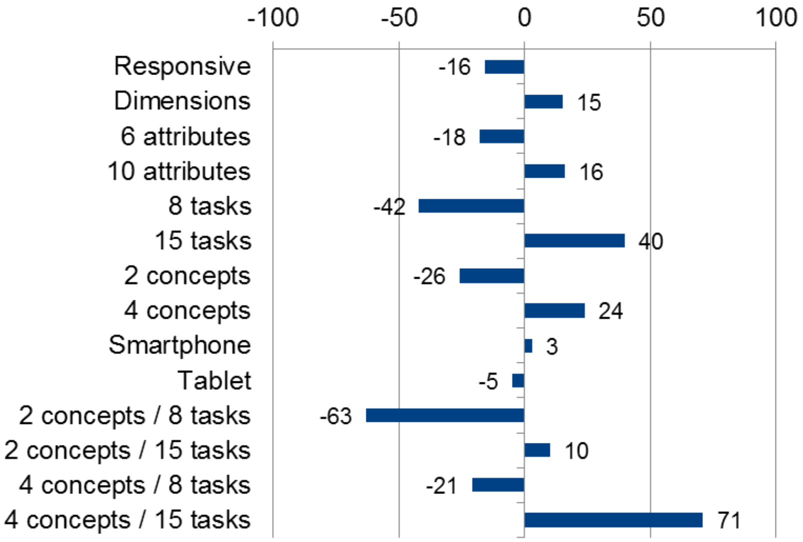
## Environment

To understand the environment in which the survey was taken, particularly for the mobile cells, questions were asked about where they took the survey and the surrounding environment they answered it in. Across the entire mobile sample, 90% of respondents took the survey at home, 5% at work with the remaining 5% in a public place. The figures differed slightly when comparing where smartphone users took the survey compared to tablet users. For smartphone users, 85% took the survey at home, 7% at work and 8% in a public place, whereas for tablet users the figures were 95%, 2% and 3% respectively. Therefore while smartphone users are more likely to be on-the-go, it appears that this is not the primary purpose for using a mobile device to answer surveys. Probably due to a higher proportion of smartphone respondents conducting the survey in public places there is a small difference in the general environment of where the survey was conducted with 4% stating that they answered the survey in a very busy or distracting environment (compared to 1% for tablet respondents). Overall, more than 80% claimed to have taken the survey in a quiet and relaxed environment which would indicate that we can reject the hypothesis that any differences we see in conjoint results are attributable to the distracting nature of the environment. One has to note though that these figures are based on claimed behaviour and not observed behaviour. Therefore, the numbers might be biased as some respondents might fear negative consequences when admitting they took the study in a distracting environment.

## Time to Complete the Conjoint Survey

The time taken to complete the conjoint part of the survey was captured to test the hypothesis that the Angular JS platform is quicker, as well as to test for differences in time taken to complete equivalent designs on a mobile device. Across all mobile cells the grand mean time taken to complete the conjoint element was 192 seconds. Figure 5 shows the difference in time taken between the grand mean and the different experimental conditions.

**Figure 5. Average Interview Time (Sec), Deviation from Grand Mean**



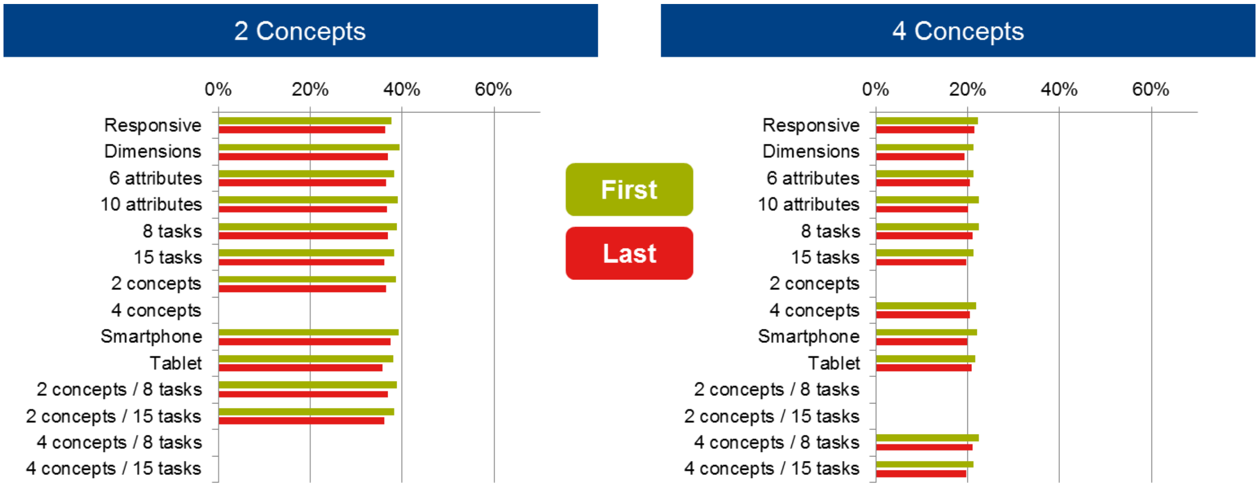
Across all cells that used the Angular JS platform the average time to complete the conjoint section was 176 seconds. This compares to 207 seconds for the cells that used the Dimensions platform, which represents an 18% increase in time and thus proving the hypothesis that this type of responsive design is quicker than the standard Dimensions platform. Increasing the number of tasks from 8 to 15 increased the time needed by 55% (150s vs. 232s) and doubling the number of concepts from 2 to 4 concepts increased the time needed by 30% (166s vs. 216s). Interestingly there is very little difference in time taken for those that answered on smartphone versus a tablet despite the need to do extra scrolling to review the text.

It was also possible to directly compare timings between equivalent designs conducted on a mobile device (using Dimensions) and a fixed device. For the simple 6 attribute, 8 task, 2 concept design it took on average 128 seconds to complete the conjoint experiment on a mobile device compared to 131 seconds on the fixed device, while for the complex design (10 attributes, 15 tasks, 4 concepts) the fixed cell was slightly quicker, averaging 298 seconds compared to 311 seconds. Previous studies (Kurz *et al.* 2016) have indicated that conjoint exercises taken on a mobile device take longer and this research shows similar results when the design is complex, albeit only marginally, but timings are comparable when designs are much simpler.

## Behavioural Differences

Due to the smaller screen size when conducting a mobile survey there is a hypothesis that there may be inherent biases regarding the position of the concept that is being chosen. Figure 6 shows the percentage of times the first and last concept is chosen across the mobile cells (results split by the 2-concept cells and the 4-concepts cells). Across each of the experimental conditions and device types there appears to be a very marginal bias towards the first concept being chosen more often. Typically the difference is less than two percentage points for the 2 concept design and less than 1.5 percentage points for the 4 concept design. When comparing positional bias of the mobile cells against the fixed cells, there is no such bias in the fixed designs towards the first concept.

**Figure 6. Percent of Times the First and Last Concept Was Chosen**



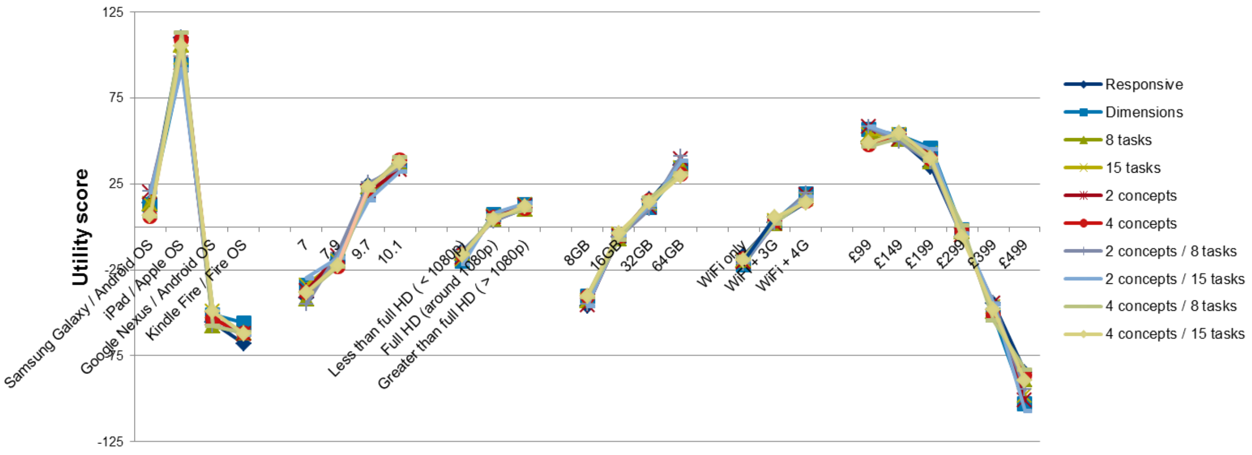
Other behavioural analysis between the mobile and fixed cells was also conducted which evaluated the proportion of times that the None option was selected for each task, and across all tasks how often the most frequently chosen concept position was chosen (flat-lining) and whether there were any behavioural changes in terms of selecting the None option more often later in the tasks. The results were very consistent across both mobile and fixed cells and results are in line with what would be expected given the design conditions.

Analysis was also conducted to identify the number of reversals in the individual level part-worth utility scores. Across the attributes which had an a priori order, for the low complexity design there were on average 4.8 reversals in the mobile cells compared to an average of 4 reversals for the fixed cells. The figures for the high complexity design were 9.5 and 8.1 respectively. In reviewing the number of reversals by attribute, the difference seen above are purely a result of differences in the price attribute as the number of reversals were very similar between device types for all other attributes.

## Utility Structure

Figure 7 shows the average re-scaled (zero-centered) part-worth utilities for the six attributes that were included across all designs. Attributes relating to Brand, Screen size, Screen resolution, Storage and Connectivity have almost identical part-worth utility structures across the different experimental conditions. For Price there are a number of reversals between the £99 and £149 level. Within the UK market the entry level Apple tablet retails from c.£249 whereas the Kindle Fire is generally available from £99. As no prohibitions were included in the design, during the conjoint experiment there would have been numerous occasions where an Apple product was shown at a price significantly lower than is seen in the real-world. Given that the majority of respondents in the mobile cells had a tablet and therefore would have a good knowledge of the pricing structure of the market it is believed that these reversals are a cause of psychological pricing and the belief that products at the lower end of the price range tested would have doubts about the quality of the product. This is likely to explain when the number of reversals in the mobile cells is higher than the fixed cells. It is also interesting to note that the reversals tend to be where the 4 concept experimental condition is being used. Overall the correlation between the part-worth utility parameters across the 18 cells was over 0.92.

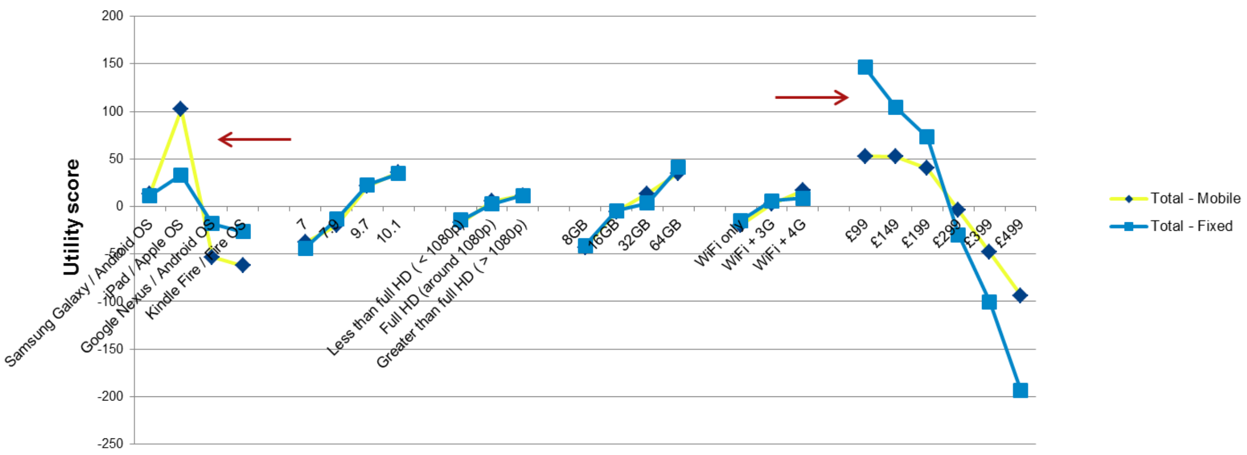
## Figure 7. Part-Worth Utility Scores by Experimental Condition



The biggest difference in utility structure that was identified was the difference between the mobile cells and the fixed cells. There are significant differences in the utility structure for the Brand and Price attributes. Within the mobile sample, there is a large increase in part-worth utility for the Apple iPad, at the expense of the Google Nexus and Kindle Fire. Mobile cells are also less price sensitive especially at the lower end of the price range tested. This again suggests a high level of market knowledge in this category and the pricing of tablets. The part-worth utility structure is flat between £99–£199 and only after this price does the elasticity become more pronounced. The fixed cells are less likely than the mobile cells to have a tablet and they show very high levels of price elasticity across all price points. The comparison of the part-worth utility structure between the mobile and fixed cells is shown in Figure 8.

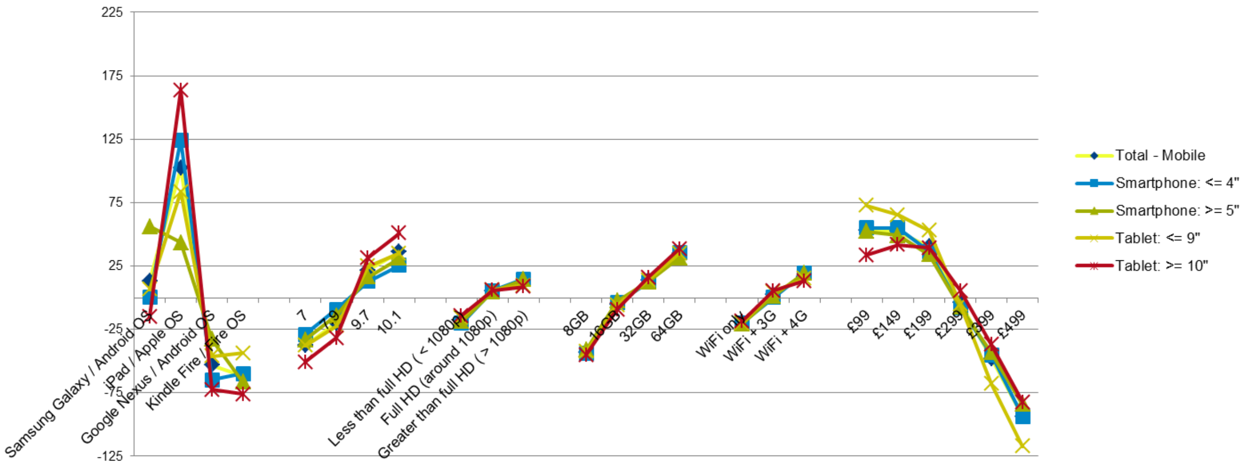
This result is not unexpected. White (2013) also found the same pattern, albeit less pronounced than in this study. He hypothesized that this is a result of the study area and that respondents answering on a mobile device have a strong affinity to brands that they own. To test this, he conducted a second study in a non-technology area and found no significant differences in utility structure.

## Figure 8. Part-Worth Utility Scores by Mobile and Fixed Cells



There is also strong evidence from this research to back up this hypothesis. Analysis of the part-worth utility structures split by device and screen size showed that respondents answering the survey on a smartphone with a screen size greater than 5" had a very different set of part-worth utilities for Brand. Unlike the other cells, rather than the Apple iPad being the most preferred, for this group of respondents the Samsung Galaxy was most preferred. At the time of the survey (March 2015), Samsung was the most dominant brand in the 5"+ market (as the iPhone 6 had not been fully launched) and almost half the sample in the smartphone/5"+ cell answered the survey on a Samsung device. Similarly, from Figure 9, the Apple iPad part-worth utility is most dominant in the Tablet 10"+ cell, and this is where Apple is dominant over Samsung with more than 75% of respondents in this cell answering the survey on an Apple iPad. Apple is much less dominant in the < 9" tablet market and this cell (38% answered on an Apple iPad and 15% on a Kindle Fire) show a much lower part-worth utility for the Apple iPad while the Kindle Fire displays a higher preference than in the other cells. It is also noted that the main reversal in the price attribute comes from the tablet 10"+ sample which further enhances the hypothesis of psychological pricing issues as the pricing of these products typically start from £300.

**Figure 9. Part-Worth Utility Scores by Device/Screen Size**



## Predictive Validity

For this research, tablet sales data had been obtained so an external validation was conducted to simulate the sales of the top 20 tablet products. The attributes/levels had been designed based on the sales data so that it was possible to do this even for cells that contained 6 attributes. For the purposes of the simulation, cells with 6 attributes had a utility structure of 0 for the levels within the 4 attributes that were not included.

However, the results were not satisfactory as the average MAE's were very high. When comparing simulated results against real-world shares there are many external factors which are not taken in to account such that when conducting simulations there is the possibility that real-world shares do not align well to simulated shares. These factors can be extensive and include awareness, distributions, marketing campaigns, stock availability, effect of sales force, etc. Two of the leading products are the Apple iPad Air 16GB and the Galaxy Tab 3. The Apple iPad has double the sales of the Samsung Galaxy despite the latter product having a better technical specification and being almost 50% cheaper, so when simulating shares using the part-worth utilities the Samsung Galaxy tablet obtained a significantly higher share than the Apple iPad, resulting in very large errors in the shares across these two products.

Two holdout tasks were incorporated into each design so it is possible to conduct analysis to determine internal prediction accuracy. Outputs such as hit rates, mean absolute error (MAE) and root likelihood (RLH) are typically used in these cases. For this study, in addition to running internal validation, MAE figures were also calculated based on out-of-sample analysis, for example, using the utility estimates obtained in cell 1 to predict the holdout data for cells 2–16 (only cells with the same number of attributes were analysed in the out-of-sample analysis) and so on.

While this output can provide useful diagnostics of internal validity they can also be misleading. For example, it is not possible to directly compare a 2 concept design with a 4 concept design as it would be expected that hit rates and MAE would be lower for the 4 concept designs. It is also common among some bodies when comparing MAE's across cells to multiply the part-worth utilities by an exponent factor in order to minimise MAE. This is done in an attempt to remove the effects of scale across the cells. For this study, a Swait-Louviere test was conducted on all possible combinations of experimental conditions (where feasible) and the tests indicated in most cases that any differences in preference are too strong to attribute the differences between the cells to only scale factor. As it is therefore unknown whether the exponent is adjusting for scale or actual differences in the preference structure it was also felt that any output based on this would be misleading so results are not reported.

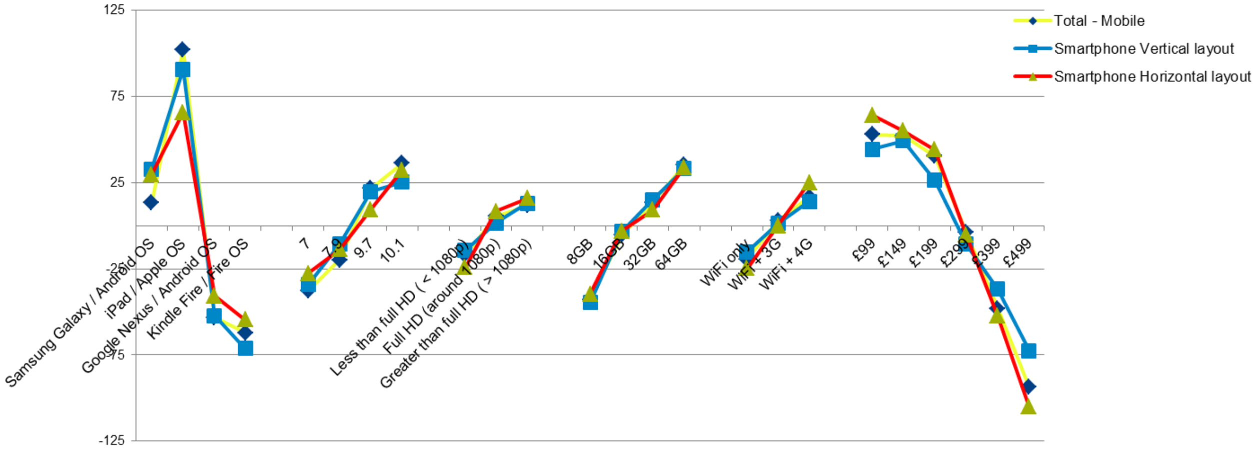
## Horizontal vs. Vertical Concept Layout

While not an initial objective of the study, additional analysis was conducted to understand whether there were any differences observed depending on whether the concepts were presented horizontally or vertically. This was possible because with the responsive Angular JS design the choice of whether the concepts appeared horizontally or vertically was automatic based on the orientation and screen size of the device being used. For respondents using a tablet, more than 98% of respondents had the tablet in a landscape orientation and as such concepts appeared horizontally. These respondents also did not change to a portrait orientation so it was not possible to do any analysis on respondents who used a tablet. However, for respondents who used a smartphone, those who went through the Angular JS platform would see the concepts in a vertical layout regardless of orientation and screen size, whereas respondents who went through the Dimensions platform would see the concepts in a horizontal layout regardless of orientation and screen size. Therefore, it is possible to compare smartphone users who went through the Angular JS platform to those who went through the Dimensions platform to understand if there are any differences between horizontal and vertical layouts.

Figure 10 shows the differences in part-worth utility scores for the 6 attributes that were present in all designs. For comparison, the part-worth utilities for all mobile respondents have been included. Reversals can be seen for the vertical layout and the magnitude (and subsequently the importance) for the Brand/Product attribute is also higher in the vertical layout.

The higher importance for the Brand/Product attribute may suggest additional simplification strategies being applied in the case of the vertical layout but when looking at behavioural diagnostics this does not appear to be the case. As discussed previously, across the mobile cells, the proportion of times the first concept was chosen was consistently higher than the proportion of times the last concept was chosen. When comparing the horizontal and vertical cells, it is the horizontal layout that produces the biggest bias with 41.1% choosing the first concept (in the 2 concept design) and 38.2% the last concept. This compares to the vertical layout which had figures of 37.3% and 36.9% respectively. A similar pattern can also be observed in the 4-concept designs where the horizontal layout had proportions of 21.4% and 18.7% and the vertical layout had proportions of 22.6% and 21.3%.

**Figure 10. Part-Worth Utility Scores by Layout**



When looking at the number of reversals by the different layouts, for the 6-attribute designs the vertical design had more reversals (5.5) than the horizontal layout (4.9). This compared to 5.1 across all mobile respondents. A similar pattern was observed in the 10-attribute designs also (Vertical = 7.5, Horizontal = 6.9). An interesting observation recorded was that the number of reversals did not differ much when comparing the 2-concepts designs against the 4-concept designs for the vertical layout (5.4 vs. 5.6 in the 6-attribute designs) but was much more pronounced for the horizontal layout (4.7 vs. 5.2).

In terms of the user experience, similar scores were recorded for both layouts although the vertical layout did score marginally higher in some of the statements (Figure 11).

**Figure 11. User Experience Scores**

| | Horrible vs Fun | Complicated vs Easy (to answer) | Difficult vs Easy (to read) | Boring vs Interesting | Willingness to use again |
|---|---|---|---|---|---|
| **Smartphone / Vertical** | 3.6 | 4.1 | 4.3 | 3.4 | 4.2 |
| **Smartphone / Horizontal** | 3.5 | 4.1 | 4.2 | 3.4 | 4.2 |

Further work would need to be conducted to look at the composition of respondents in both these cells as the research has shown that the device that was used to answer the survey and the knowledge of market prices could have caused any of the differences that have been observed. While the two cells are not directly comparable as the vertical layout was conducted within an Angular JS platform and the horizontal layout completed in a Dimensions layout there is little to suggest based on the work done that using a vertical layout will result in significantly different results than the more conventional horizontal layout.

## CONCLUSIONS/RECOMMENDATIONS

The research has shown that there are no significant issues with conducting conjoint on a mobile device as long as the overall questionnaire structure has been optimized for mobile. Tablet users did have a better user experience than smartphone users and while differences between fixed cells and mobile cells are significant, the mobile cells scored highly on the key user experience questions.

There are large demographic differences between respondents using mobile devices and those that use fixed devices so it is not recommended to conduct mobile device-only research unless there is a specific requirement to target those particular types of respondents.

The analysis has shown that respondents are able to complete conjoint designs with a high cognitive burden (4 concepts, 17 tasks, 10 attributes) with little difficulty or degradation in survey experience. While respondents could cope with 10 attribute designs which involved scrolling it should be noted that the attribute text in this research was relatively little. Only minor differences were seen in utility structure between the experimental conditions (reversals in Price utility more prominent in the 4-concept designs) and where there were differences these could be explained through psychological pricing issues or due to the brand affinity that respondents had which is a reflection of the brand of the device they used to conduct the survey (smartphone or tablet).

While there was little sign of degradation the data showed slightly better user experience for the 2 concept/15 task design compared to the 4 concept/8 task design, although the time to complete the conjoint experiment is slightly longer for the 2 concept/15 task design. Unless a responsive platform is being used a 2 concept design will require less scrolling which is an advantage over the 4 concept design.

Based on the designs used in this research, standard interviewing platforms such as Dimensions, as long as they are optimized for mobile research, appear to perform as well as more sophisticated responsive platforms. However, if designs are likely to include heavy text and/or use of images, then a responsive design is likely to be more suitable. Responsive designs are also proven to be quicker due to the single page application and generally scored higher on the user-experience ratings.

Hypotheses regarding whether it is the environment that may affect the quality of mobile user rather than the device were rejected on the basis that more than 9 in 10 conduct the survey at home and 8 in 10 do so in a quiet and relaxed environment.

The analysis conducted to understand whether there were any behavioural changes in the way mobile users answer conjoint tasks did not show any significant differences. There appears to be a very minor first concept bias (1–2 percentage points) but this is unlikely to affect results. Across the experimental conditions there was a consistent pattern towards the first concept across all conditions. There were also no differences identified in the proportion of times the None option was selected (across experimental conditions or across device type) and the increase in reversals may be explained by the mobile sample having a better knowledge of pricing in the tablet area.

In summary:

- Conducting conjoint on mobile devices, even complex designs is no problem if the questionnaire is optimized for mobile.
- Those answering on a tablet had a slightly better user experience that those answering on a smartphone but both scored highly.
- Respondents answering on a mobile device exhibit a different demographic structure than those who answer on PC/Laptop.
- There is no need to oversimplify conjoint designs for mobile. Respondents can comfortably cope with 10 attributes when text is light.
- It is still better to keep things simple so 2 concepts per task is preferred to 4.
- For complex designs that involves long text, lots of concepts and/or graphics, consider a responsive platform.
- The vertical layout did not affect results significantly.
- Studies optimized for mobile do not necessarily take longer to complete especially with a simple design.

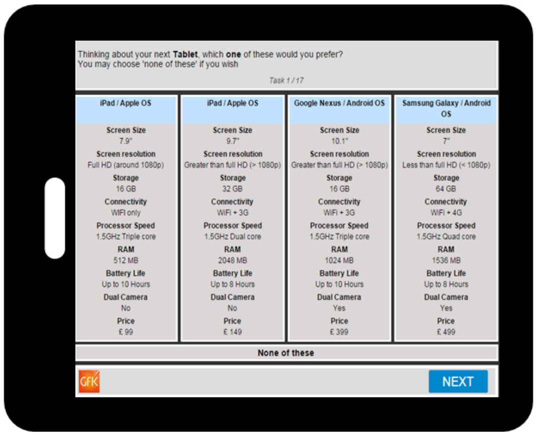
Chris Moore

Christian Neuerburg

## Appendix A—Concept Orientation

### Tablet—Responsive Design

Landscape



Thinking about your next **Tablet**, which **one** of these would you prefer?
You may choose 'none of these' if you wish

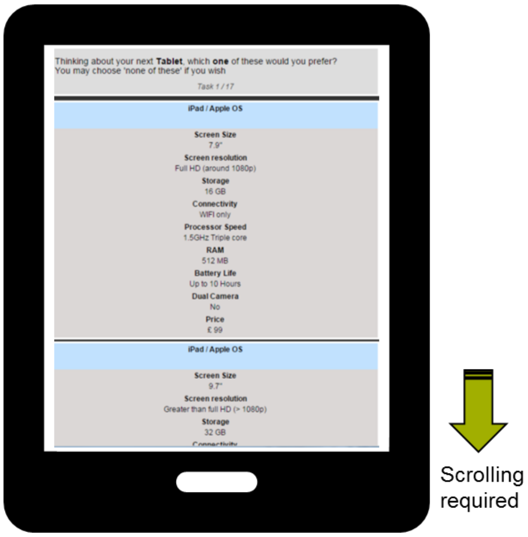
*Task 1 / 17*

| iPad / Apple OS | iPad / Apple OS | Google Nexus / Android OS | Samsung Galaxy / Android OS |
|---|---|---|---|
| **Screen Size**<br>7.9" | **Screen Size**<br>9.7" | **Screen Size**<br>10.1" | **Screen Size**<br>7" |
| **Screen resolution**<br>Full HD (around 1080p) | **Screen resolution**<br>Greater than full HD (> 1080p) | **Screen resolution**<br>Greater than full HD (> 1080p) | **Screen resolution**<br>Less than full HD (< 1080p) |
| **Storage**<br>16 GB | **Storage**<br>32 GB | **Storage**<br>16 GB | **Storage**<br>64 GB |
| **Connectivity**<br>WIFI only | **Connectivity**<br>WiFi + 3G | **Connectivity**<br>WiFi + 3G | **Connectivity**<br>WiFi + 4G |
| **Processor Speed**<br>1.5GHz Triple core | **Processor Speed**<br>1.5GHz Dual core | **Processor Speed**<br>1.5GHz Triple core | **Processor Speed**<br>1.5GHz Quad core |
| **RAM**<br>512 MB | **RAM**<br>2048 MB | **RAM**<br>1024 MB | **RAM**<br>1536 MB |
| **Battery Life**<br>Up to 10 Hours | **Battery Life**<br>Up to 8 Hours | **Battery Life**<br>Up to 10 Hours | **Battery Life**<br>Up to 8 Hours |
| **Dual Camera**<br>No | **Dual Camera**<br>No | **Dual Camera**<br>Yes | **Dual Camera**<br>Yes |
| **Price**<br>£ 99 | **Price**<br>£ 149 | **Price**<br>£ 399 | **Price**<br>£ 499 |

**None of these**

No scrolling required

Portrait



Thinking about your next **Tablet**, which **one** of these would you prefer?
You may choose 'none of these' if you wish

*Task 1 / 17*

**iPad / Apple OS**

**Screen Size**
7.9"
**Screen resolution**
Full HD (around 1080p)
**Storage**
16 GB
**Connectivity**
WIFI only
**Processor Speed**
1.5GHz Triple core
**RAM**
512 MB
**Battery Life**
Up to 10 Hours
**Dual Camera**
No
**Price**
£ 99

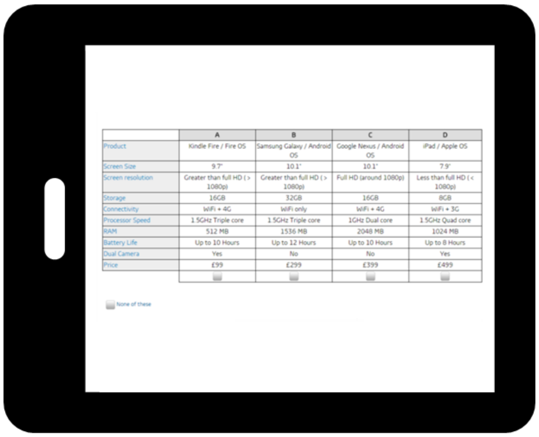
**iPad / Apple OS**

**Screen Size**
9.7"
**Screen resolution**
Greater than full HD (> 1080p)
**Storage**
32 GB
**Connectivity**

Scrolling required

## Dimensions—Responsive Design

Landscape

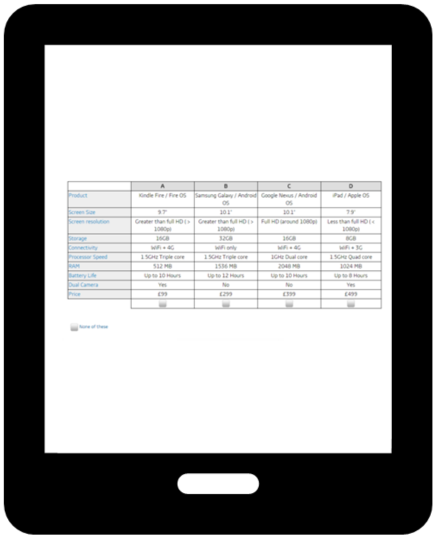


| | A | B | C | D |
|---|---|---|---|---|
| Product | Kindle Fire / Fire OS | Samsung Galaxy / Android OS | Google Nexus / Android OS | iPad / Apple OS |
| Screen Size | 9.7" | 10.1" | 10.1" | 7.9" |
| Screen resolution | Greater than full HD ( > 1080p) | Greater than full HD ( > 1080p) | Full HD (around 1080p) | Less than full HD ( < 1080p) |
| Storage | 16GB | 32GB | 16GB | 8GB |
| Connectivity | WiFi + 4G | WiFi only | WiFi + 4G | WiFi + 3G |
| Processor Speed | 1.5GHz Triple core | 1.5GHz Triple core | 1GHz Dual core | 1.5GHz Quad core |
| RAM | 512 MB | 1536 MB | 2048 MB | 1024 MB |
| Battery Life | Up to 10 Hours | Up to 12 Hours | Up to 10 Hours | Up to 8 Hours |
| Dual Camera | Yes | No | No | Yes |
| Price | £99 | £299 | £399 | £499 |
| | □ | □ | □ | □ |

□ None of these

No scrolling required

Portrait



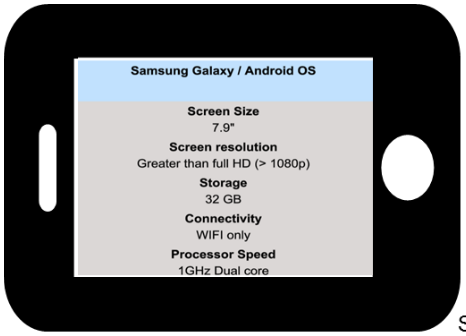
| | A | B | C | D |
|---|---|---|---|---|
| Product | Kindle Fire / Fire OS | Samsung Galaxy / Android OS | Google Nexus / Android OS | iPad / Apple OS |
| Screen Size | 9.7" | 10.1" | 10.1" | 7.9" |
| Screen resolution | Greater than full HD ( > 1080p) | Greater than full HD ( > 1080p) | Full HD (around 1080p) | Less than full HD ( < 1080p) |
| Storage | 16GB | 32GB | 16GB | 8GB |
| Connectivity | WiFi + 4G | WiFi only | WiFi + 4G | WiFi + 3G |
| Processor Speed | 1.5GHz Triple core | 1.5GHz Triple core | 1GHz Dual core | 1.5GHz Quad core |
| RAM | 512 MB | 1536 MB | 2048 MB | 1024 MB |
| Battery Life | Up to 10 Hours | Up to 12 Hours | Up to 10 Hours | Up to 8 Hours |
| Dual Camera | Yes | No | No | Yes |
| Price | £99 | £299 | £399 | £499 |
| | □ | □ | □ | □ |

□ None of these

No scrolling required

## Smartphone—Responsive Design

Landscape



**Samsung Galaxy / Android OS**

**Screen Size**
7.9"

**Screen resolution**
Greater than full HD (> 1080p)

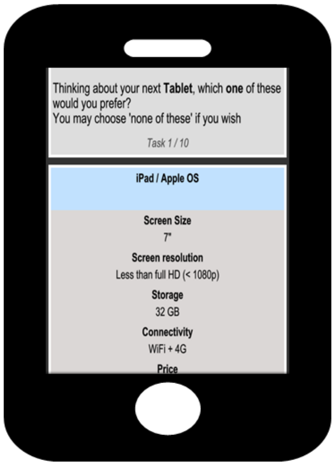
**Storage**
32 GB

**Connectivity**
WIFI only

**Processor Speed**
1GHz Dual core

Scrolling required

Portrait



Thinking about your next **Tablet**, which **one** of these would you prefer?
You may choose 'none of these' if you wish

*Task 1 / 10*

**iPad / Apple OS**

**Screen Size**
7"

**Screen resolution**
Less than full HD (< 1080p)

**Storage**
32 GB

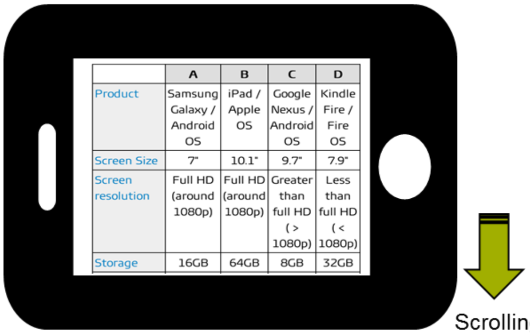
**Connectivity**
WiFi + 4G

**Price**

Scrolling required

## Smartphone—Dimensions

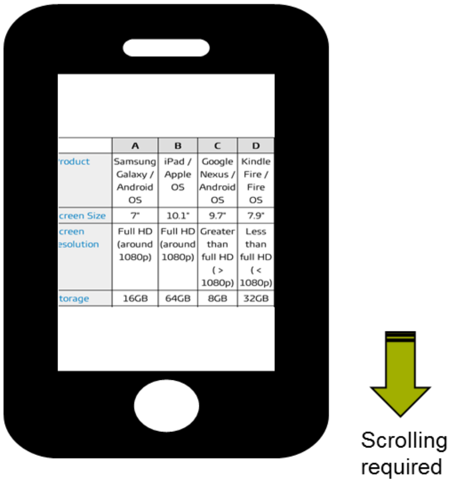
Landscape



|  | A | B | C | D |
|---|---|---|---|---|
| Product | Samsung Galaxy / Android OS | iPad / Apple OS | Google Nexus / Android OS | Kindle Fire / Fire OS |
| Screen Size | 7" | 10.1" | 9.7" | 7.9" |
| Screen resolution | Full HD (around 1080p) | Full HD (around 1080p) | Greater than full HD ( > 1080p) | Less than full HD ( < 1080p) |
| Storage | 16GB | 64GB | 8GB | 32GB |

Scrolling required

Portrait



|  | A | B | C | D |
|---|---|---|---|---|
| roduct | Samsung Galaxy / Android OS | iPad / Apple OS | Google Nexus / Android OS | Kindle Fire / Fire OS |
| creen Size | 7" | 10.1" | 9.7" | 7.9" |
| creen esolution | Full HD (around 1080p) | Full HD (around 1080p) | Greater than full HD ( > 1080p) | Less than full HD ( < 1080p) |
| torage | 16GB | 64GB | 8GB | 32GB |

Scrolling required

## Appendix B—Conjoint Attributes

|  | Attributes | Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Level 6 |
|---|---|---|---|---|---|---|---|
| 1 | Product | Samsung Galaxy / Android OS | iPad / Apple OS | Google Nexus / Android OS | Kindle Fire / Fire OS |  |  |
| 2 | Screen Size | 7" | 7.9" | 9.7" | 10.1" |  |  |
| 3 | Screen resolution | Less than full HD ( < 1080p) | Full HD (around 1080p) | Greater than full HD ( > 1080p) |  |  |  |
| 4 | Storage | 8GB | 16GB | 32GB | 64GB |  |  |
| 5 | Connectivity | WiFi only | WiFi + 3G | WiFi + 4G |  |  |  |
| 6 | Processor Speed | 1GHz Dual core | 1.5GHz Dual core | 1.5GHz Triple core | 1.5GHz Quad core |  |  |
| 7 | RAM | 512 MB | 1024 MB | 1536 MB | 2048 MB |  |  |
| 8 | Battery Life | Up to 8 Hours | Up to 10 Hours | Up to 12 Hours |  |  |  |
| 9 | Dual Camera | Yes | No |  |  |  |  |
| 10 | Price | £99 | £149 | £199 | £299 | £399 | £499 |

Attributes highlighted in blue formed the 6 attribute design

## References

Diener, Chris *et al.* (2013), "Making Conjoint Mobile: Adapting Conjoint to the Mobile Phenomenon," Sawtooth Software Proceedings 2013, pg 55–68.

Kurz, Peter *et al.*, "Smartphones vs. Desktop—Discrete Choice Models on Mobile Devices," SKIM/Sawtooth European Conference 2016.

Swait, Joffre & Louviere, Jordan (1993), "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," Journal of Marketing Research 1993, pg 305–314.

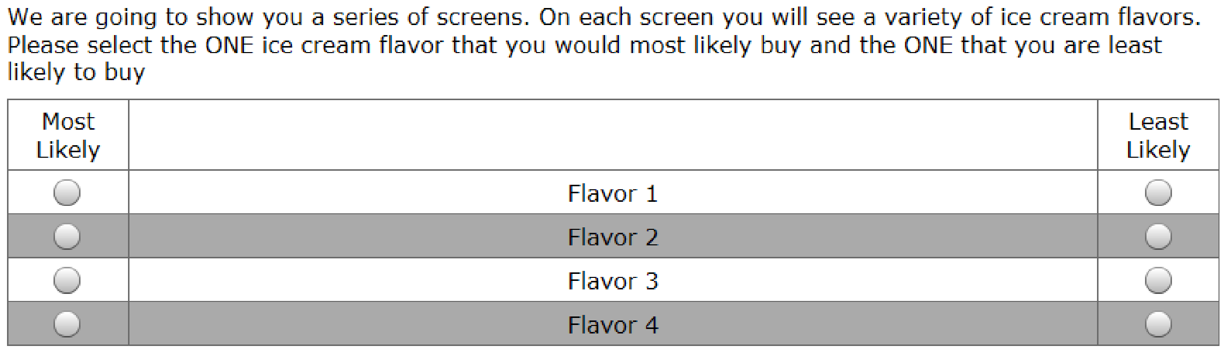
White, Joseph (2013), "Choice Experiments in Mobile Web Environment," Sawtooth Software Proceedings 2013, pg 69–82.

# Can Adaptive MaxDiff Provide Better Results than Standard MaxDiff?

**Howard Firestone**
*RTi Research*

## Background

Maximum Difference Scaling (a.k.a. MaxDiff) is a choice-based tradeoff technique that is widely used to understand the value of members of a list of items. Respondents are shown a series of screens (typically comprised of 3 to 5 items) and asked to select the "best" and "worst" item on each screen. Here is an example of a MaxDiff screen.



We are going to show you a series of screens. On each screen you will see a variety of ice cream flavors. Please select the ONE ice cream flavor that you would most likely buy and the ONE that you are least likely to buy

| Most Likely | | Least Likely |
|---|---|---|
| ○ | Flavor 1 | ○ |
| ○ | Flavor 2 | ○ |
| ○ | Flavor 3 | ○ |
| ○ | Flavor 4 | ○ |

Historically, MaxDiff has provided a hierarchy of the *relative* appeal or importance of the items being measured.

We, at RTi, report this hierarchy as indices (where the average equals 100). We calculate the indices from the rescaled scores (or the probability sheet from Lighthouse Studio files) by simply multiplying the scores by the number of items on the list. In the study we will be referencing in this article, we included 28 flavors in the MaxDiff exercise—therefore, we multiplied the Average MaxDiff scores by 28 to calculate the unanchored indices.

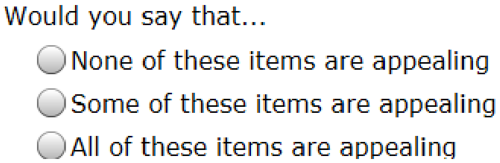
| Label | Average | Average X 28 |
|---|---|---|
| Flavor 1 | 3.00 | 84 |
| Flavor 2 | 5.41 | 151 |
| . . . | | |
| Flavor 28 | 5.15 | 144 |
| Sum | 100.00 | |
| Average | | 100 |

About 8–10 years ago, both Jordan Louviere (Indirect Dual Response method) and Kevin Lattery (Direct Binary method) developed anchoring techniques that added an *absolute* dimension to the MaxDiff scores. Prior to the development of these anchoring techniques, you wouldn't know if the most appealing items were truly appealing or if they were the best of a weak set of items.

Both of these anchoring techniques compare the utilities to a “reference point” or an “average item.” When the utilities are more appealing than the reference point, the indices are above 100, whereas items that are less appealing have indices below 100.

Brief descriptions of both anchoring approaches follow:

- **Indirect Dual Response Method**—On each MaxDiff screen, respondents are asked to select all/some or none of the items.



Would you say that…

○ None of these items are appealing
○ Some of these items are appealing
○ All of these items are appealing

  We don’t use this technique because we would expect a response of “Some of these items are appealing” to most MaxDiff screens. With this response, we learn how only 2 (of the 4 or 5) items are rated relative to the reference.

- **Direct Binary Method**—With this technique, a follow-up question is asked after the last MaxDiff screen.

  *Below are some ice cream flavors that you have evaluated. Which, if any, of the following flavors would you be <u>likely to purchase</u>? (**SELECT <u>ALL</u> THAT APPLY)***

  We use this question on most studies and typically report it as an index. Depending on the appeal or importance of the items included in the MaxDiff exercises, we find that the average of all indices can be higher or lower than 100.

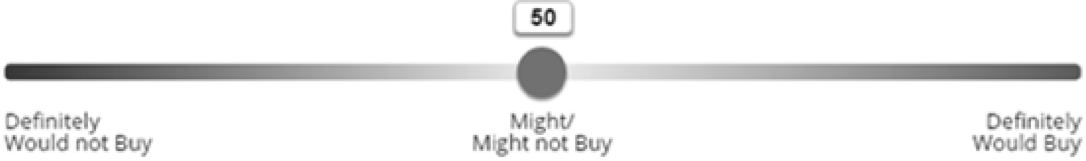
  We typically include all items in the follow-up question. Sawtooth Software’s MaxDiff scores on-the-fly functionality can also be used to select a subset of items for the follow-up.

While we have successfully employed the Direct Binary Method on numerous studies, a few concerns have surfaced over time:

1. Indices that don’t average to 100 are counter-intuitive to clients who are accustomed to indices that average 100.
2. The wording of the follow-up question can influence the response to the anchoring question and impact results. For example, should the anchor be asked as “Which of these flavors are appealing” vs. “Which of these flavors are *very* appealing.”
3. When using MaxDiff scores on-the-fly functionality, Sawtooth Software recommends showing each item 4 times in order to obtain reasonable individual level score estimation for follow-up questioning. Since we typically include each item on 2 MaxDiff screens we usually include all items in the follow-up anchor question.

As a result of these concerns, we sought an alternative approach—an Adaptive MaxDiff technique where we changed both the composition of the items shown in the MaxDiff exercise and the follow-up exercise.

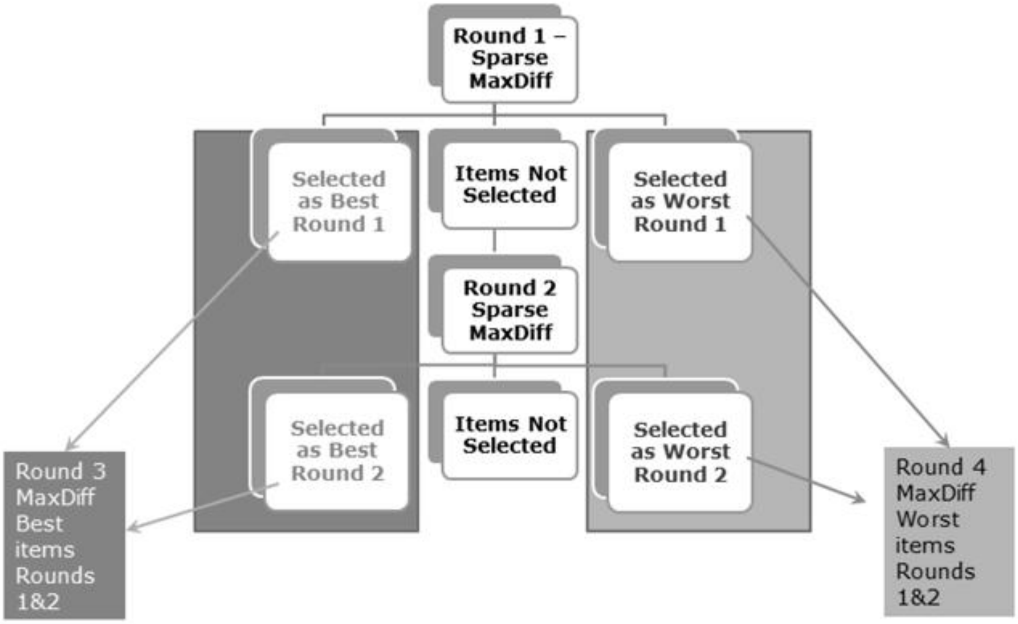
- For the MaxDiff exercise, we felt that we would get better individual level discrimination between items by including:
    - MaxDiff tasks comprised of the respondent's "preferred" ("Best") items only. The items that are the "Best of the Best" are then included in the follow-up slider question described below.
    - Other MaxDiff tasks that contained only the respondent's "least preferred" ("Worst") items. The items that are the "Worst of the Worst" are also included in the follow-up slider question.
    - Other MaxDiff tasks comprised of items that were not selected as either best or worst.
- For the follow-up question, we replaced the Direct Binary Method with a 100 point slider which enables us to convert the HB scores to a 100 point scale. This scale provides a better gauge of the appeal or importance of each item because it provides an end point at the upper end of the scale. The follow-up slider question was only asked of a few items (the "best of the best" and "worst of the worst") which shortened the respondent exercise.



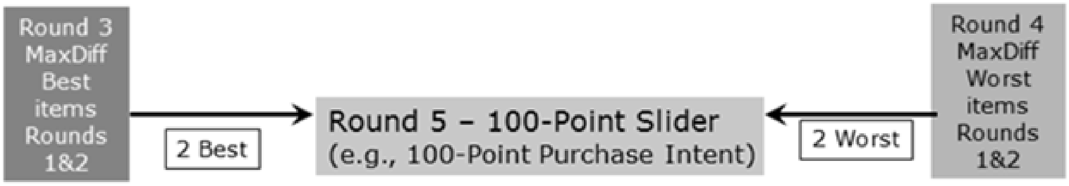
The Adaptive MaxDiff approach was comprised of 4 rounds of MaxDiff Screens.

- Round 1—Sparse MaxDiff (where each item appears on only 1 screen) of all the items. This round is used to assign items to 3 groups—**Preferred/Not selected/Less Preferred.**
- Round 2—Sparse MaxDiff of the items **"Not Selected"** in Round 1—The **Not Selected** group tends to be the largest group from Round 1. This round provides additional learning of preferences from this large number of items. Furthermore, the items selected as "Best" are added to the **Preferred** group and the items selected as "Worst" are added to **Less Preferre**d. This increases the likelihood that a "preferred" item is placed in the **"Preferred"** item group and was not mis-assigned because it was on a screen with another "preferred" item.
- Round 3—MaxDiff of the **Preferred** ("Best") items selected in Rounds 1 and 2. This round develops the hierarchy of the preferred items. It also identifies the "Best of the Best" which will serve as the "High" end of the respondent's scale when the HB scores are converted to the 100 point scale.
- Round 4—MaxDiff of the **Less Preferred** ("Worst") items selected in Rounds 1 and 2. This round identifies the "Worst of the Worst" which will serve as the "Low" end of the respondents scale when the respondent's HB scores are converted to the 100 point scale.

The schematic of the approach follows.



As mentioned earlier, the follow-up exercise was a 100 point slider for the items that are the "Best of the Best" from Round 3 and the items that are the "Worst of the Worst" from round 4.


Round 3 MaxDiff Best items Rounds 1&2
2 Best
Round 5 – 100-Point Slider
(e.g., 100-Point Purchase Intent)
2 Worst
Round 4 MaxDiff Worst items Rounds 1&2

While the primary focus of this paper is to illustrate the value of the slider follow-up, we also assessed the viability of using Lattery's Direct Binary anchoring question among a subset of the items.

- For the proof of concept study, we asked the Direct Binary question for the items that were not selected from the first 2 rounds of MaxDiff screens.
- Alternatively, we could have selected a mix of items from the MaxDiff exercise.

We expected the Adaptive MaxDiff approach to have the following advantages over standard MaxDiff approaches:

- Better individual level discrimination between items
- Easier understanding of reporting by replacing indices with a 100 scaling framework
- Shorter MaxDiff exercise for the respondent

## ADAPTIVE MAXDIFF—PROOF OF CONCEPT STUDY METHODOLOGY

The next section of this paper describes the research approach we employed to assess the Adaptive concept.

The study we conducted was comprised of 1,000 adults who had purchased and eaten ice cream in the past 6 months. Each respondent was assigned to 1 of 6 cells.

- 2 cells used Standard MaxDiff exercise
  - They differed in terms of the number of MaxDiff screens and the number of times each item was seen
  - Both cells included Lattery's Direct Binary follow-up among all flavors
- 2 cells employed Adaptive MaxDiff
  - 1 used the slider follow-up
  - The other used Lattery's Direct Binary Anchor among a subset of flavors for the follow-up
- 2 cells of holdout tasks

The methodological comparison between the 4 MaxDiff cells is provided in the table below.

| | **Cell 1 Standard MaxDiff** | **Cell 2 Standard MaxDiff** | **Cell 3 Adaptive MaxDiff** | **Cell 4 Adaptive MaxDiff** |
|---|---|---|---|---|
| # MaxDiff Screens | 17 | 12 | 14 | 14 |
| # Items / MaxDiff Screen | 5 | 5 | 4 /5 | 4/5 |
| Follow-up Question | Direct Binary Anchor—all 28 flavors | Direct Binary Anchor—all 28 flavors | Direct Binary Anchor—8 flavors | 100 Point Slider—4 flavors (Best of Best and Worst of Worst) |

More detailed descriptions of the 6 cells follows.

- 4 MaxDiff cells—all cells included 28 ice cream flavors/200 respondents per cell
  - Cell 1—Standard MaxDiff exercise in which each item is shown on 3 MaxDiff screens
    - 17 MaxDiff screens—5 items per screen
    - Lattery's Direct Binary Anchor asked of all 28 flavors spread across 3 screens
  - Cell 2—Standard MaxDiff exercise in which each item is shown on 2 MaxDiff screens
    - 12 MaxDiff screens—5 items per screen
    - Direct Binary Anchor asked of all 28 flavors spread across 3 screens
  - Cell 3—Adaptive MaxDiff with Direct Binary Anchor
    - 4 rounds of MaxDiff screens (total of 14 screens)
    - Round 1—Sparse design of all 28 flavors—6 screens, 4 had 5 items/2 had 4 items
    - Round 2—Sparse design of items not selected in Round 1—4 screens, 4 items each
    - Round 3—Winners from rounds 1 and 2—2 screens of 5 items each

        - Round 4—Losers from rounds 1 and 2—2 screens of 5 items each
        - Follow-up question—Direct Binary anchor of the 8 items not selected in either Round 1 or Round 2
    - Cell 4—Adaptive MaxDiff with Slider Follow-up
        - 4 rounds of MaxDiff screens—Same as Cell 3
        - Follow-up question—100 point slider for 4 items
            - 2 items selected as best in round 3
            - 2 items selected as worst in round 4
- 2 Holdout Task cells—100 respondents per task
    - Holdout cell 1—2 holdout tasks
        - Ranking of 8 of the 28 flavors
        - Purchase interest of 8 other flavors
    - Holdout cell 2—2 holdout tasks
        - Ranking of a different set of 8 flavors
        - Purchase interest of a different set of 8 flavors

The grid below describes the MaxDiff design that was used for the 2 Adaptive cells.

| Round 1 | | Round 2 | | Round 3 | | Round 4 | |
|---|---|---|---|---|---|---|---|
| Set 1 | Item 1 | Set 7 | S1 NS-1 | Set 11 | S8 Winner | Set 13 | S2 Loser |
| | Item 2 | | S2 NS-1 | | S3 Winner | | S9 Loser |
| | Item 3 | | S5 NS-1 | | S1 Winner | | S10 Loser |
| | Item 4 | | S4 NS-1 | | S10 Winner | | S5 Loser |
| | Item 5 | | | | S5 Winner | | S3 Loser |
| Set 2 | Item 6 | Set 8 | S3 NS-1 | Set 12 | S7 Winner | Set 14 | S1 Loser |
| | Item 7 | | S5 NS-2 | | S2 Winner | | S8 Loser |
| | Item 8 | | S4 NS-2 | | S6 Winner | | S4 Loser |
| | Item 9 | | S6 NS-1 | | S4 Winner | | S7 Loser |
| | Item 10 | | | | S9 Winner | | S6 Loser |
| Set 3 | Item 11 | Set 9 | S2 NS-2 | | | | |
| | Item 12 | | S3 NS-2 | | | | |
| | Item 13 | | S4 NS-3 | | | | |
| | Item 14 | | S1 NS-2 | | | | |
| Set 4 | Item 15 | Set 10 | S1 NS-3 | | | | |
| | Item 16 | | S6 NS-2 | | | | |
| | Item 17 | | S2 NS-3 | | | | |
| | Item 18 | | S5 NS-3 | | | | |
| | Item 19 | | | | | | |
| Set 5 | Item 20 | | | | | | |
| | Item 21 | | | | | | |
| | Item 22 | | | | | | |
| | Item 23 | | | | | | |
| | Item 24 | | | | | | |
| Set 6 | Item 25 | | | | | | |
| | Item 26 | | | | | | |
| | Item 27 | | | | | | |
| | Item 28 | | | | | | |

One hundred versions of the Round 1 design were generated using Sawtooth Software's MaxDiff designer with the following specifications:

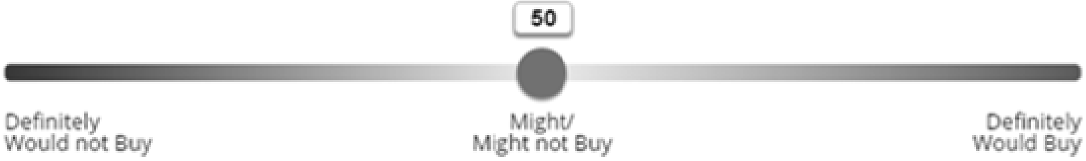
- Number of Items (Attributes) = 28
- Number of Items per Set (Question) = 4
- Number of Sets (Questions) per Respondent = 6
- Allow Individual Designs Lacking Connectivity = Yes

This accounted for 24 of the 28 flavors in each version. For each version, we scrambled the 4 items that were not included and assigned them to Sets 1, 2, 4 and 5. This means that 4 screens contained 5 items and the other 2 screens contained 4 items.

Round 2 was structured so that each set was comprised of items that came from different sets in Round 1. Each set in round 3 included 3 winners from Round 1, and 2 winners from Round 2 while each set in Round 4 contained 3 losers from Round 1, and 2 losers from Round 2.

## CELL 4—ADAPTIVE MAXDIFF WITH THE SLIDER FOLLOW-UP

The next section of this paper describes how the slider scores are processed in HB and how scores are converted to the 100 point scale.



In Cell 4 we asked respondents to use a 100 point slider to tell us how interested they would be in purchasing 4 of the ice cream flavors that had been included in the MaxDiff exercise. Each respondent rated the 2 flavors that had been selected "best" in Round 3 (MaxDiff screens from the items selected as "best" in earlier rounds) and the 2 flavors that were rated "worst" in Round 4 (MaxDiff screens from items selected as "worst" in Rounds 1 and 2).

Each of the slider scores were compared to an anchor such that the sum of the slider score plus the anchor would equal 100. Using this suggestion, there would be 4 additional tasks—1 for each of the 4 slider scores as follows.

| | Best 1 | Best 2 | Worst 1 | Worst 2 |
|---|---|---|---|---|
| Flavor | 1 | 2 | 3 | 28 |
| Slider Score | 90 | 70 | 40 | 25 |
| Value for Choice Task (Slider Score/100) | 0.90 | 0.70 | 0.40 | 0.25 |
| Anchor (100-Slider Score) | 10 | 30 | 60 | 75 |
| Value for Choice Task (Anchor/100) | 0.10 | 0.30 | 0.60 | 0.75 |

Using CBC/HB, the tasks would be coded as follows:

| | | Flavor | Flavor | Flavor | | Flavor | |
|---|---|---|---|---|---|---|---|
| | **Task** | **1** | **2** | **3** | **. . .** | **28** | **Choice** |
| **Best 1** | 1 | 1 | 0 | 0 | | 0 | 0.90 |
| | 1 | 0 | 0 | 0 | | 0 | 0.10 |
| **Best 2** | 2 | 0 | 0 | 0 | | 0 | 0.70 |
| | 2 | 0 | 0 | 0 | | 0 | 0.30 |
| **Worst 1** | 3 | 0 | 0 | 0 | | 0 | 0.40 |
| | 3 | 0 | 0 | 0 | | 0 | 0.60 |
| **Worst 2** | 4 | 0 | 0 | 0 | | 0 | 0.25 |
| | 4 | 0 | 0 | 0 | | 0 | 0.75 |

Note that the anchor is coded as 0s for all flavors. In addition to creating the appropriate relationship between the slider scores when generating the HB scores, this anchoring technique helps to establish 50 as the slider score for an average item. Thus a score of 50 on a 100 point slider would be equivalent to an index of 100 when reporting anchored MaxDiff scores.

Since the purchase interest slider was obtained for only 4 flavors, we used the following process to interpolate purchase intent scores for the other 24 flavors for each respondent.

1. Calculate the difference between the highest and lowest HB score for each respondent
   HB Range = HB Max minus HB Min
   - Since we had included the slider scores as additional tasks, we knew that the item that received the higher slider score would have the highest HB and the item with the lower slider score would have the lowest HB score
2. Calculate the difference between the highest and lowest slider scores for each respondent
   Slider Range = Slider Max minus Slider Min
3. For each respondent, calculate the difference between each item's HB score and the lowest HB score
   Item HB diff = HB Item minus HB min
4. Slider scores for each respondent were interpolated using the following formula:
   Item Slider score = (Item HB diff / HB Range * Slider Range) + Slider Min

An example of the interpolation calculation follows:

| | **Example Scores** |
|---|---|
| HB Score | 3.72 |
| HB Max | 6.21 |
| HB Min | -6.05 |
| Slider Max | 90 |
| Slider Min | 15 |

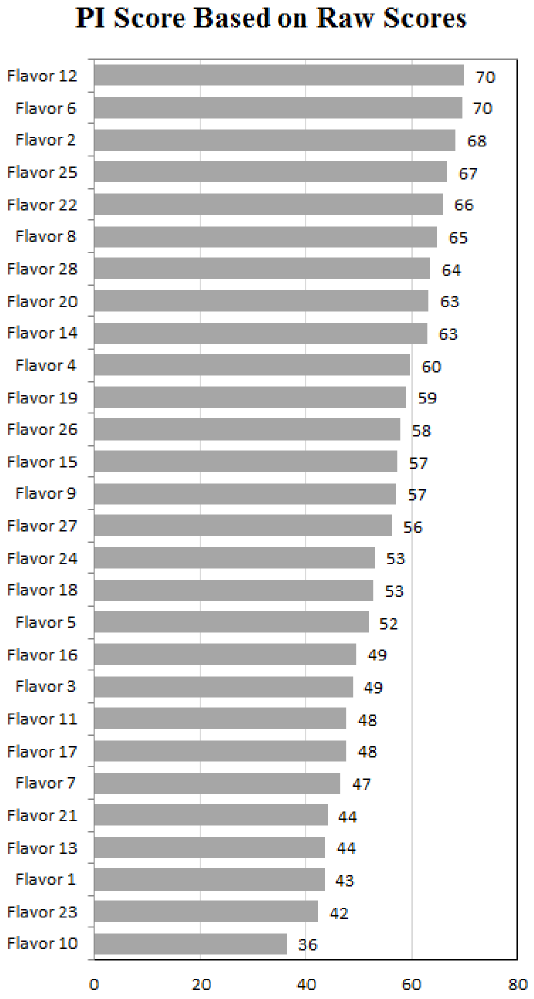
| | Calculation | Result |
|---|---|---|
| HB Range (HB Max - HB Min) | 6.21 minus -6.05 | 12.26 |
| Slider Range (Slider Max - Slider Min) | 90 minus 15 | 75 |
| Item HB Diff (HB Score - HB Min) | 3.72 minus -6.05 | 9.77 |
| Item Slider Score | [(9.77 / 12/26) * 75] + 15 | |
| | 9.77 / 12.26 | 79.70% |
| | 79.7% * 75 | 59.80 |
| | 59.8 + 15 | 74.80 |

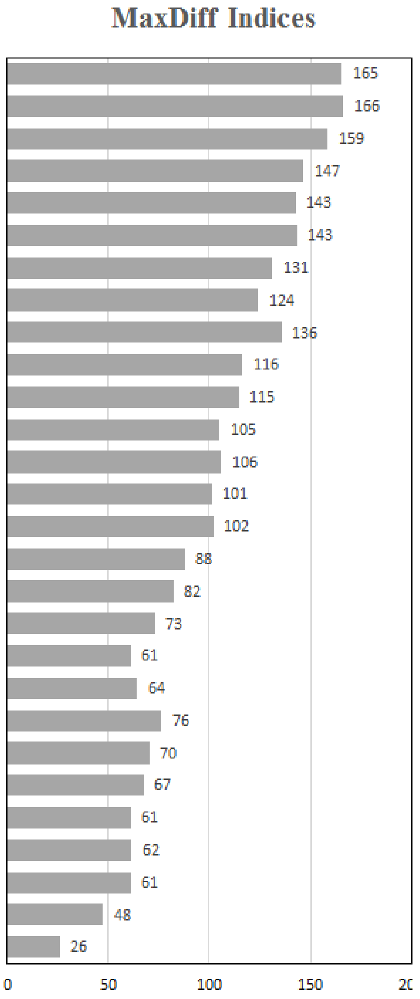
For this example, the item Slider Score = 74.80

## Key Learnings from Proof of Concept Study

### Key Finding #1—Converting scores to a 100 point client-friendly scale generates similar findings to the MaxDiff Indices.

The following exhibit compares the Interpolated Purchase Intent scores to the Unanchored MaxDiff Indices for Cell 4 and shows that the hierarchy of results is very similar as reflected by the correlation 0.99 for the actual scores.

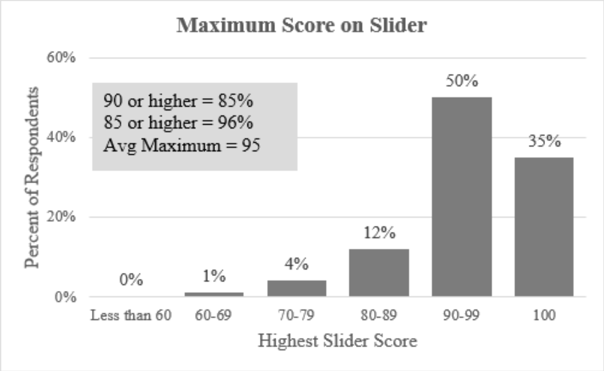


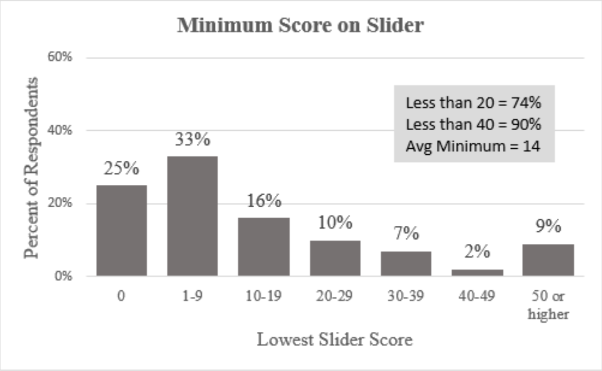


In the exhibit above, a Purchase Intent score of 55 would be equivalent to an unanchored MaxDiff Index of 100. As you can see from the following analyses, this equals the midpoint between the Average Maximum (95) and Average Minimum (14) slider scores.

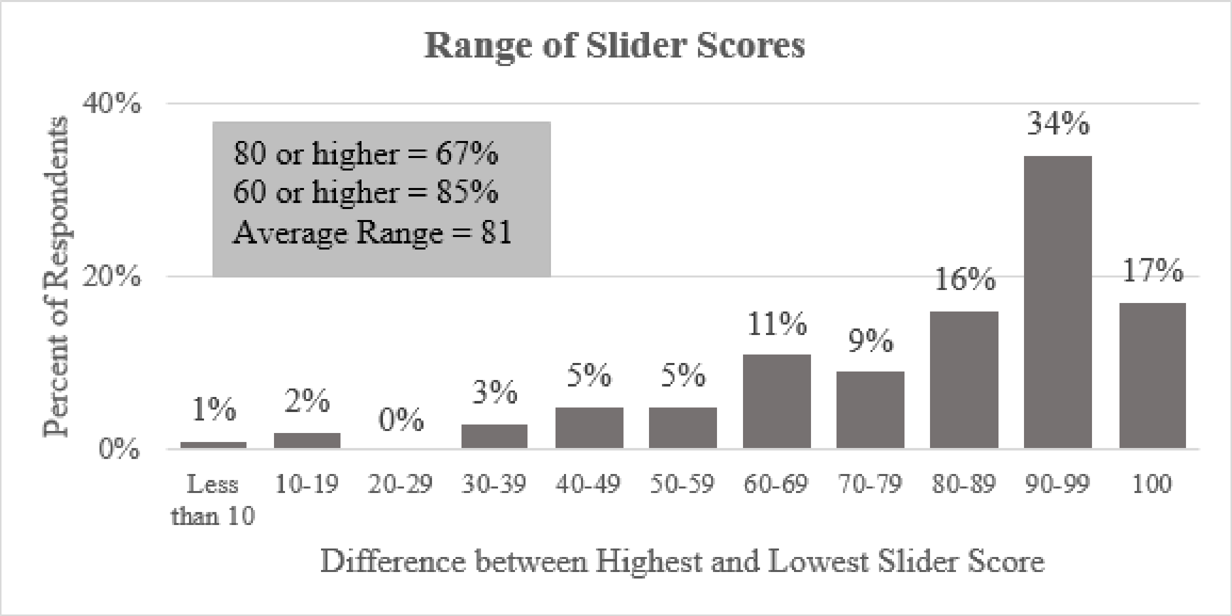
### Key Finding #2—The range of slider scores helps to convey strong individual respondent discrimination between items.

In cell 4, the Adaptive MaxDiff with the slider follow-up, most respondents gave very high purchase intent scores to the flavors they preferred and very low scores to the flavors they liked least suggesting very good discrimination between these flavors.



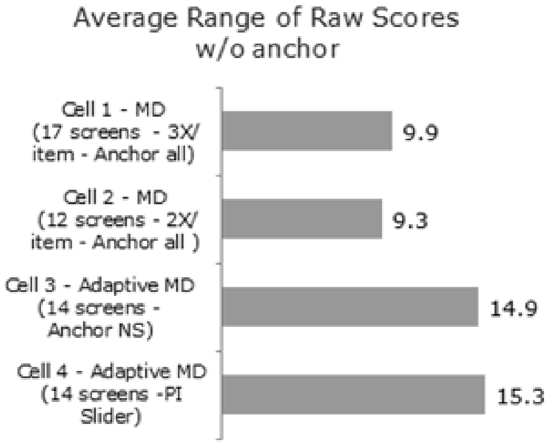


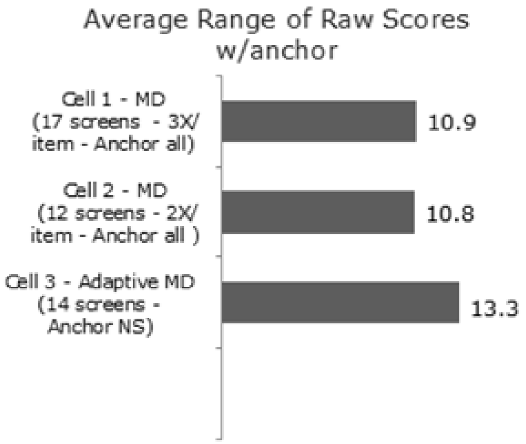
When comparing the maximum and minimum scores, most respondents had a very wide range of slider scores and only 2 respondents gave essentially the same slider scores (as indicated by a range of less than 10) which would be equivalent to straight-lining when using a rating scale.



## Key Finding #3—Individual respondent discrimination between flavors was higher for the 2 Adaptive MaxDiff cells.

As mentioned earlier, we expected individual level results to be more discriminating for the 2 Adaptive approaches than they were for the Standard approaches. This expectation was borne out when we looked at the range of both unanchored and anchored HB scores for individual respondents.


Average Range of Raw Scores w/o anchor

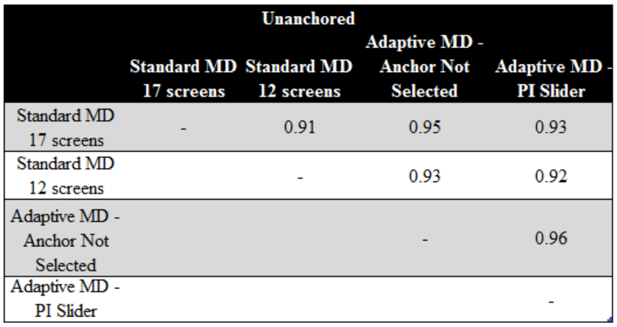

Average Range of Raw Scores w/anchor

## Key Finding #4—The results from the Adaptive MaxDiff cells were similar to the Standard MaxDiff cells.

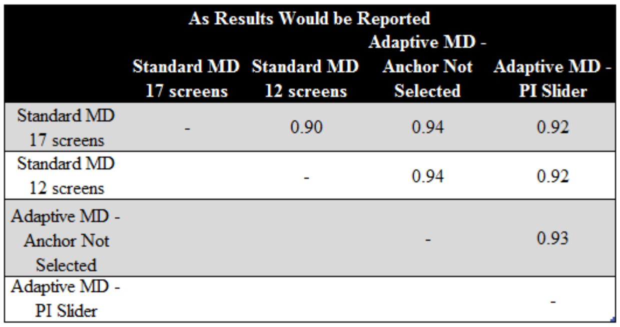
The pattern of results were very similar across the 4 cells as noted by the high correlations between the 4 cells when looking at both the Indices from the Unanchored HB scores and the results as they would be reported. For both of the Standard MaxDiff cells and 1 of the 2 Adaptive MaxDiff cells, the scores would be reported as Indices from the anchored HB scores, while the Adaptive cell that includes the slider would be reported as Averages on a 100 point scale. These results are included in the appendix as:

- Appendix Table 1—Unanchored MaxDiff Indices
- Appendix Table 2—Results as they Would be Reported

The pattern of results was very similar for both analyses—as between cell correlations are very high.



| | Unanchored | | | |
|---|---|---|---|---|
| | Standard MD 17 screens | Standard MD 12 screens | Adaptive MD - Anchor Not Selected | Adaptive MD - PI Slider |
| Standard MD 17 screens | - | 0.91 | 0.95 | 0.93 |
| Standard MD 12 screens | | - | 0.93 | 0.92 |
| Adaptive MD - Anchor Not Selected | | | - | 0.96 |
| Adaptive MD - PI Slider | | | | - |



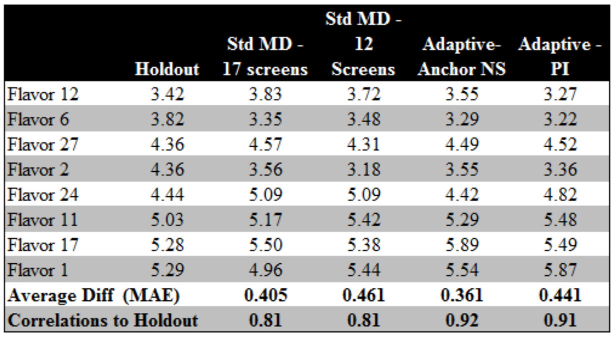
| | As Results Would be Reported | | | |
|---|---|---|---|---|
| | Standard MD 17 screens | Standard MD 12 screens | Adaptive MD - Anchor Not Selected | Adaptive MD - PI Slider |
| Standard MD 17 screens | - | 0.90 | 0.94 | 0.92 |
| Standard MD 12 screens | | - | 0.94 | 0.92 |
| Adaptive MD - Anchor Not Selected | | | - | 0.93 |
| Adaptive MD - PI Slider | | | | - |

### Key Finding #5—The results from the Adaptive and Standard MaxDiff Approaches were similar when compared to Holdout Tasks

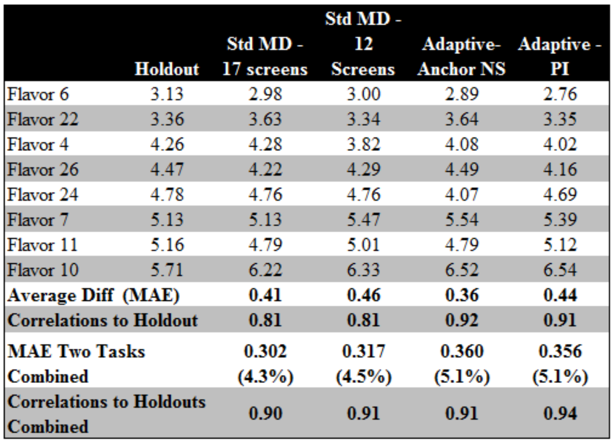
The results from the 4 cells were compared to 4 out-of-sample holdout tasks.

1. 2 holdout tasks were comprised of Rankings of 8 flavors.
2. The other 2 holdout tasks were purchase intent scores using a standard 5 point purchase intent scale.

For the holdout tasks where flavors were ranked, the analysis was based on the Average ranking of the 8 flavors. For all 4 cells the pattern of rankings are similar to the holdout tasks as the correlations were strong when these 2 holdout tasks are combined. The Average difference (MAE) tends to be slightly higher for the 2 Adaptive cells than they are for the 2 Standard cells.

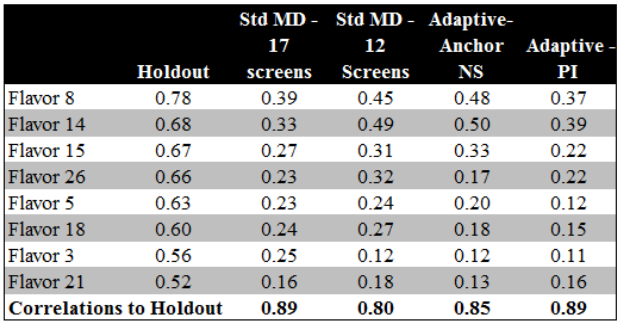


| | Holdout | Std MD - 17 screens | Std MD - 12 Screens | Adaptive-Anchor NS | Adaptive - PI |
|---|---|---|---|---|---|
| Flavor 12 | 3.42 | 3.83 | 3.72 | 3.55 | 3.27 |
| Flavor 6 | 3.82 | 3.35 | 3.48 | 3.29 | 3.22 |
| Flavor 27 | 4.36 | 4.57 | 4.31 | 4.49 | 4.52 |
| Flavor 2 | 4.36 | 3.56 | 3.18 | 3.55 | 3.36 |
| Flavor 24 | 4.44 | 5.09 | 5.09 | 4.42 | 4.82 |
| Flavor 11 | 5.03 | 5.17 | 5.42 | 5.29 | 5.48 |
| Flavor 17 | 5.28 | 5.50 | 5.38 | 5.89 | 5.49 |
| Flavor 1 | 5.29 | 4.96 | 5.44 | 5.54 | 5.87 |
| **Average Diff (MAE)** | | **0.405** | **0.461** | **0.361** | **0.441** |
| **Correlations to Holdout** | | **0.81** | **0.81** | **0.92** | **0.91** |

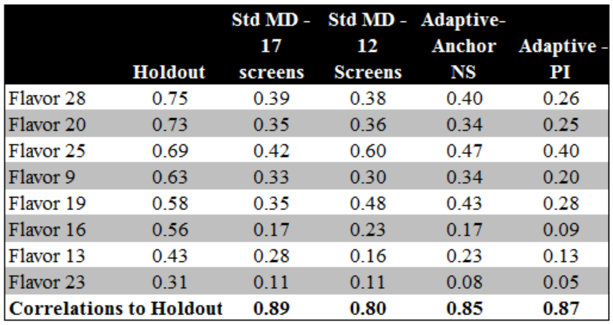


| | Holdout | Std MD - 17 screens | Std MD - 12 Screens | Adaptive-Anchor NS | Adaptive - PI |
|---|---|---|---|---|---|
| Flavor 6 | 3.13 | 2.98 | 3.00 | 2.89 | 2.76 |
| Flavor 22 | 3.36 | 3.63 | 3.34 | 3.64 | 3.35 |
| Flavor 4 | 4.26 | 4.28 | 3.82 | 4.08 | 4.02 |
| Flavor 26 | 4.47 | 4.22 | 4.29 | 4.49 | 4.16 |
| Flavor 24 | 4.78 | 4.76 | 4.76 | 4.07 | 4.69 |
| Flavor 7 | 5.13 | 5.13 | 5.47 | 5.54 | 5.39 |
| Flavor 11 | 5.16 | 4.79 | 5.01 | 4.79 | 5.12 |
| Flavor 10 | 5.71 | 6.22 | 6.33 | 6.52 | 6.54 |
| **Average Diff (MAE)** | | **0.41** | **0.46** | **0.36** | **0.44** |
| **Correlations to Holdout** | | **0.81** | **0.81** | **0.92** | **0.91** |
| **MAE Two Tasks Combined** | | **0.302**<br>**(4.3%)** | **0.317**<br>**(4.5%)** | **0.360**<br>**(5.1%)** | **0.356**<br>**(5.1%)** |
| **Correlations to Holdouts Combined** | | **0.90** | **0.91** | **0.91** | **0.94** |

When the results from these 4 cells were compared to the 2 holdouts using a standard purchase intent question, the pattern of responses was again similar to the holdout tasks (as reflected by the fairly strong correlations). However, both the MaxDiff Indices (for cells 1 thru 3) and the Purchase Intent scores estimated (for cell 4) are substantially lower than the Top Box Purchase Intent levels for the holdout tasks. This likely reflects the overstatement inherent with rating scales rather than any concerns with the MaxDiff exercise.



| | Holdout | Std MD - 17 screens | Std MD - 12 Screens | Adaptive-Anchor NS | Adaptive - PI |
|---|---|---|---|---|---|
| Flavor 8 | 0.78 | 0.39 | 0.45 | 0.48 | 0.37 |
| Flavor 14 | 0.68 | 0.33 | 0.49 | 0.50 | 0.39 |
| Flavor 15 | 0.67 | 0.27 | 0.31 | 0.33 | 0.22 |
| Flavor 26 | 0.66 | 0.23 | 0.32 | 0.17 | 0.22 |
| Flavor 5 | 0.63 | 0.23 | 0.24 | 0.20 | 0.12 |
| Flavor 18 | 0.60 | 0.24 | 0.27 | 0.18 | 0.15 |
| Flavor 3 | 0.56 | 0.25 | 0.12 | 0.12 | 0.11 |
| Flavor 21 | 0.52 | 0.16 | 0.18 | 0.13 | 0.16 |
| **Correlations to Holdout** | | **0.89** | **0.80** | **0.85** | **0.89** |



| | Holdout | Std MD - 17 screens | Std MD - 12 Screens | Adaptive-Anchor NS | Adaptive - PI |
|---|---|---|---|---|---|
| Flavor 28 | 0.75 | 0.39 | 0.38 | 0.40 | 0.26 |
| Flavor 20 | 0.73 | 0.35 | 0.36 | 0.34 | 0.25 |
| Flavor 25 | 0.69 | 0.42 | 0.60 | 0.47 | 0.40 |
| Flavor 9 | 0.63 | 0.33 | 0.30 | 0.34 | 0.20 |
| Flavor 19 | 0.58 | 0.35 | 0.48 | 0.43 | 0.28 |
| Flavor 16 | 0.56 | 0.17 | 0.23 | 0.17 | 0.09 |
| Flavor 13 | 0.43 | 0.28 | 0.16 | 0.23 | 0.13 |
| Flavor 23 | 0.31 | 0.11 | 0.11 | 0.08 | 0.05 |
| **Correlations to Holdout** | | **0.89** | **0.80** | **0.85** | **0.87** |

## CONCLUSIONS

We recommend using Adaptive MaxDiff with the Slider follow-up because it has a number of advantages compared to a Standard MaxDiff exercise.

- The primary advantage of the Adaptive MaxDiff with Slider follow-up is that the reporting framework is more client-friendly and intuitive as Indices (that do not average 100) are replaced by a straightforward 100 point scale.
    - Furthermore it provides a gauge of how well an item performs relative to a top end value of 100.
    - A score of 50 is average which is comparable to an index of 100 for an Anchored MaxDiff exercise.
    - In addition to reporting the scores as averages, you can report the percent of respondents who have interpolated scores that meet or exceed a threshold.
- Item hierarchies for individual respondents are clearer as reflected by the wider ranges of HB scores.
- Respondents who should be considered to be pulled can be identified while the study is in field, thereby improving the quality of the data, rather than waiting for the fieldwork to be completed so the "Goodness of Fit" statistic can be reviewed when the HB scores are generated.
    - Respondents would be pulled if:
        - They assign a higher slider score to a "Worst of Worst" item than they give to a "Best of Best" item.
        - The difference between the slider score for a "best" item and a "worst" is below a threshold level may suggest that the respondent is "straight-lining."
- Shortens the MaxDiff exercise for the respondent as only a few items are included in the follow-up slider question.
- The use of the slider with the endpoints and midpoint labeled reduces the concern we had regarding the question wording for the follow-up question when using Direct Binary anchor.
- Contains strong correlations to the holdout tasks.

There are, however a few disadvantages to using this approach:

- Requires more programming expertise and data preparation time (If using Sawtooth Software, the analysis must be done in CBC/HB rather than Lighthouse).
- MAE was slightly higher than the Standard MaxDiff cells.
- The Goodness of Fit statistic does not apply.

The Adaptive MaxDiff approach with Lattery's Direct Binary anchor follow-up is also a viable approach that offers advantages relative to a standard MaxDiff:

- Item hierarchies for individual respondents are clearer as the ranges of the HB scores are wider.
- Shortens the MaxDiff exercise for respondents by reducing the number of items included in the anchor follow-up.
    - Further investigation is needed to determine if the anchoring question should include a mix of preferred, not selected and less preferred items.
- Contains strong correlations to the holdout tasks.

There are, however, some disadvantages to this approach.

- Requires more programming expertise and data preparation time (If using Sawtooth Software, the analysis must be done in CBC/HB rather than Lighthouse).

- MAE was slightly higher than the Standard MaxDiff cells.
- The Goodness of Fit statistic does not apply.

## Additional Considerations

It is important to note there are a number of differences between the Adaptive MaxDiff approach and a Standard MaxDiff.

When using an Adaptive MaxDiff,

- Counts analysis should not be used.
- The design is not orthogonal even though the Sparse design created for Round 1 is “near” orthogonal.
- You cannot prohibit pairs of item from appearing together.
- Item pairs will appear together on 2 MaxDiff tasks for some respondents.
- The number of items may vary between screens.
    - The design can be modified slightly so that the same number of items can be shown on all screens**.** In our design, Round 1 had 6 MaxDiff screens—4 screens had 5 items; and 2 screens contained only 4 items. We could have taken items that were not selected on earlier screens and added them to later screens in the same round.

## Future Considerations

We plan to add a MaxDiff screen to future studies that use the Adaptive technique. This additional screen addresses the concern that all of the items in one of the Round 1 screens may be among the more preferred items, which places a preferred item in the less preferred group. This screen would include:

- The items selected as “Worst of the Best” from Round 3 (Round 3 contained MaxDiff tasks of the “Preferred items” from Rounds 1 and 2),
- The items selected as “Best of the Worst” from Round 4 (Round 4 contained MaxDiff tasks of the “Less Preferred” items from Rounds 1 and 2),
- Item(s) not selected in Rounds 1 and 2.

Howard Firestone

**Appendix Table 1. Unanchored MaxDiff Scores**

| | Std MD 17 screens | Std MD 14 Screens | Adaptive Anchor NS | Adaptive PI |
|---|---|---|---|---|
| Flavor 2 | 151 | 176 | 155 | 162 |
| Flavor 25 | 151 | 195 | 157 | 150 |
| Flavor 8 | 146 | 150 | 158 | 146 |
| Flavor 6 | 145 | 149 | 160 | 170 |
| Flavor 28 | 144 | 131 | 141 | 134 |
| Flavor 12 | 130 | 127 | 148 | 169 |
| Flavor 20 | 126 | 110 | 118 | 126 |
| Flavor 19 | 125 | 150 | 129 | 117 |
| Flavor 22 | 123 | 133 | 133 | 145 |
| Flavor 9 | 121 | 93 | 115 | 102 |
| Flavor 14 | 116 | 145 | 148 | 138 |
| Flavor 27 | 102 | 111 | 108 | 103 |
| Flavor 15 | 99 | 96 | 113 | 107 |
| Flavor 4 | 96 | 120 | 107 | 118 |
| Flavor 26 | 93 | 98 | 97 | 107 |
| Flavor 1 | 90 | 69 | 78 | 61 |
| Flavor 11 | 87 | 82 | 92 | 77 |
| Flavor 18 | 87 | 82 | 76 | 83 |
| Flavor 5 | 85 | 81 | 76 | 74 |
| Flavor 13 | 84 | 47 | 67 | 62 |
| Flavor 3 | 83 | 52 | 61 | 64 |
| Flavor 24 | 79 | 74 | 110 | 89 |
| Flavor 7 | 71 | 59 | 62 | 68 |
| Flavor 17 | 68 | 68 | 56 | 71 |
| Flavor 21 | 61 | 54 | 52 | 61 |
| Flavor 16 | 55 | 65 | 65 | 61 |
| Flavor 23 | 54 | 49 | 32 | 47 |
| Flavor 10 | 29 | 33 | 26 | 25 |

**Appendix Table 2. As Results Would Be Reported**

| | Std MD 17 screens | Std MD 12 Screens | Adaptive Anchor NS | Adaptive PI |
|---|---|---|---|---|
| Flavor 6 | 191 | 211 | 210 | 70 |
| Flavor 25 | 189 | 261 | 201 | 67 |
| Flavor 2 | 188 | 229 | 193 | 68 |
| Flavor 8 | 185 | 203 | 212 | 65 |
| Flavor 28 | 179 | 175 | 185 | 64 |
| Flavor 12 | 173 | 195 | 194 | 70 |
| Flavor 22 | 158 | 181 | 180 | 66 |
| Flavor 19 | 157 | 210 | 180 | 59 |
| Flavor 14 | 156 | 218 | 198 | 63 |
| Flavor 20 | 156 | 156 | 148 | 63 |
| Flavor 9 | 146 | 131 | 151 | 57 |
| Flavor 27 | 135 | 166 | 144 | 56 |
| Flavor 4 | 127 | 165 | 145 | 60 |
| Flavor 26 | 120 | 138 | 113 | 58 |
| Flavor 13 | 119 | 73 | 100 | 44 |
| Flavor 15 | 119 | 133 | 143 | 57 |
| Flavor 1 | 117 | 86 | 104 | 43 |
| Flavor 11 | 112 | 121 | 124 | 48 |
| Flavor 3 | 111 | 68 | 76 | 49 |
| Flavor 18 | 107 | 118 | 94 | 53 |
| Flavor 5 | 105 | 113 | 97 | 52 |
| Flavor 24 | 99 | 98 | 137 | 53 |
| Flavor 7 | 94 | 85 | 75 | 47 |
| Flavor 17 | 90 | 95 | 80 | 48 |
| Flavor 21 | 78 | 81 | 68 | 44 |
| Flavor 16 | 77 | 99 | 86 | 49 |
| Flavor 23 | 63 | 61 | 48 | 42 |
| Flavor 10 | 33 | 46 | 43 | 36 |

# Comparing Two Methods to Estimate Missing Maximum Difference Utilities

*Kelsey White*
*Paul Johnson*
*SSI*[1]

Since the development of the Maximum Difference technique, researchers have been pushing the limits of how many items can be tested in a single exercise. What might have started as 20 to 30 items in one exercise has quickly ballooned to 40 or even 400. Having every respondent see every item quickly creates fatigue, even with Sparse MaxDiff where the respondents only see each item once instead of multiple times.

The trend of wanting more data out of fewer questions is not unique to MaxDiff. Many researchers have been looking at ways to get more data out of shorter surveys using modularization. The biggest problem with modularization is how to handle the missing data. Ralph Wirth used one technique called Express MaxDiff (Wirth & Wolfrath 2009), which systematically selects which items respondents see. This technique relies on the HB algorithm to borrow information across respondents to estimate the individual-level utilities for items a respondent did not see. We blend this technique with elements of another technique applied by Kevin Lattery (2007). Lattery uses the EM algorithm to estimate unseen tasks in Choice-Based Conjoint. We use the same technique, not to estimate the tasks themselves, but to replace the estimation of the individual-level utilities from the HB algorithm when the item was never shown to the respondent.

To test the EM algorithm against HB estimation, we select 200 policy statements made by 2016 presidential candidate Donald Trump and design an Express MaxDiff exercise. Then we use the Direct Binary method to anchor to two different thresholds; one threshold being increased likelihood of voting for Mr. Trump and one threshold being decreased likelihood of voting for Mr. Trump. We then perform several types of analyses and compare the results of the HB-only utilities with the HB+EM utilities.

## Introduction

Market researchers in today's environment face two major challenges: clients who want more data from a single survey, and respondents who want a shorter, more engaging survey experience. Discrete choice data collected via a conjoint or MaxDiff exercise requires question repetition, naturally limiting the number of attributes that can be tested if respondent fatigue is to be avoided.

In considering MaxDiff specifically, it is recommended to show each item 2–3 times to each respondent (Orme 2005). The number of tasks required for stable individual level utility

[1] Survey Sampling International

estimates scales linearly with a large item set, contributing to long interviews and causing respondent fatigue.

In an effort to avoid burdening respondents with a large number of MaxDiff tasks while still obtaining utilities for a large set of attributes, researchers have tried several methods of dealing with the issue. These methods include adaptive designs (Orme 2006), aggregate models such as the Bandit Model (Fairchild, Orme & Schwartz 2015), estimating unseen tasks (Lattery 2007), conducting an initial sorting exercise (Hendrix & Drucker 2007), showing attributes fewer times than recommended, such as Sparse MaxDiff (Wirth & Wolfrath 2009), and not showing all attributes to all respondents via an Express MaxDiff (Wirth & Wolfrath 2009).

With an interest in eliminating the tradeoff between respondent fatigue and complete data for a larger number of attributes, we decided to test the capabilities of the EM algorithm (expectation maximization) in imputing "missing" data in a prohibitively long set of MaxDiff items.

## Background

We decided to use an extreme number of items, 200 in total, in designing our MaxDiff test. We wanted to find an equally unique subject matter, and decided upon policy statements made by 2016 presidential candidate Donald Trump. We hoped to not only test utility imputation methods, but to also illustrate the usefulness of the MaxDiff technique when applied to political studies.

Mr. Trump has spent time in the political sphere as both a conservative and a liberal, providing a wide range of statements from both sides of the aisle for use in our study. Trump is also a polarizing figure with strong and simple statements that are easy for respondents to evaluate and then agree or disagree with.

We chose roughly 10 statements in each of 22 topical categories, with statements representing both conservative and liberal stances on the issues. These statements, with the categories included shown in Table 1, ranged from specific policy evaluations such as "I support the ban on assault weapons" and "I support a slightly longer waiting period to purchase a gun" to more general commentary on the current state of America, such as "Our country needs a truly great leader" and "we need a truly great leader now."

**Table 1. Statement Categories**

2nd Amendment
Corporate America
Criminal Justice Reform
Deficit/Spending
Discrimination
Education
Eminent Domain
Environment
Foreign Policy
Free speech/Media
Healthcare

Immigration
Jobs/Economy
Military/VA
Other candidates/politicians
Patriotic Statements
Religion
Social Issues
Social Programs
Taxes
Trade
Women

## Research Design/Methodology

We fielded a 20-question survey (excluding MaxDiff exercises) focused on political opinions and feelings about the media among Survey Sampling International's Online U.S. Voter Panel. We obtained n=1,500 completes from late March to early April 2016. The survey began by collecting respondents' party affiliation, recent voting patterns, and general attitude about the direction the country is headed. Quotas were in place to balance respondents on key metrics aligned with the registered U.S. voting population, including age, gender, and party affiliation, displayed below in Table 2. After the data were collected, some minor weighting was put in place to balance on ethnicity.

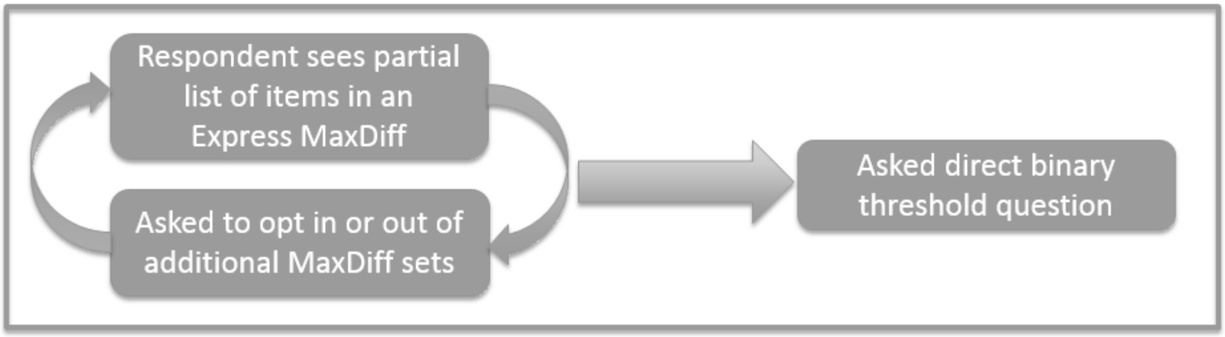
**Table 2. Respondent Quotas**

| Gender | Age | Party Affiliation |
|---|---|---|
| Male 47% | 18–34 23% | Republican 26% |
| Female 53% | 35–54 34% | Democrat 30% |
|  | 55+ 43% | Independent 44% |

After assessing respondent familiarity with current 2016 presidential candidates from the two major political parties (at the time including Hillary Clinton, Donald Trump, Bernie Sanders, Ted Cruz, and John Kasich) and collecting the candidate each respondent felt most likely to vote for at that moment in time, we began the MaxDiff portion of the study.

Figure 1, displayed below, shows the experimental design of the study. Respondents were shown a partial subset of the total 200 items in an Express MaxDiff and were then given the option to opt in to additional rounds of MaxDiff questions. Once respondents no longer opted in to additional rounds, they were shown a Direct Binary threshold question for use in creating two anchors.

**Figure 1. Survey Flow**


Respondent sees partial list of items in an Express MaxDiff
Asked to opt in or out of additional MaxDiff sets
Asked direct binary threshold question

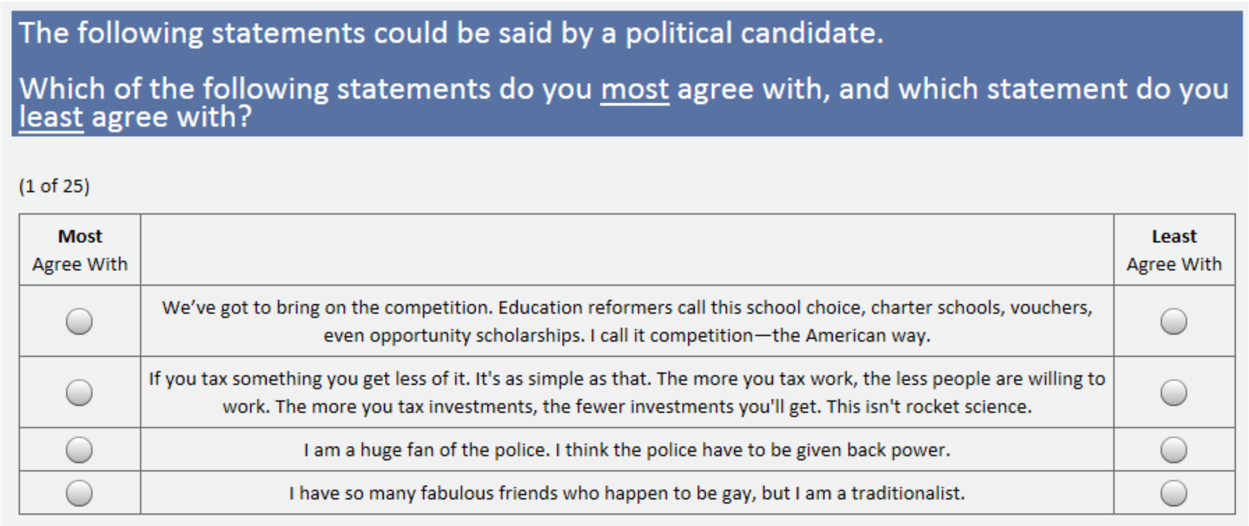
The MaxDiff portion of the study was set up as 7 MaxDiff "modules," each containing a subset of the total 200 items. The first MaxDiff module consisted of 10 fixed items and 20 additional items selected randomly, for a total of 30 items tested. Each respondent saw 25 screens with 4 statements per screen. We wanted a large amount of data on these items to ensure robust individual-level utilities for each statement seen. Table 3 displays the 10 statements chosen as fixed to be seen by all 1,500 respondents.

**Table 3. Fixed Items in the First MaxDiff Module**

| Category | Statement |
|---|---|
| Education | We've got to bring on the competition. Education reformers call this school choice, charter schools, vouchers, even opportunity scholarships. I call it competition—the American way. |
| Immigration | We have to have a wall. We have to have a border. And in that wall we're going to have a big fat door where people can come into the country, but they have to come in legally. |
| Labor Unions | Unions are about the only political force reminding us to remember the American working family. |
| Military/VA | Militarily, we're going to build up our military. We're going to have such a strong military that nobody, nobody is going to mess with us. We're not going to have to use it. |
| Other candidates/politicians | One of the key problems today is that politics is such a disgrace. Good people don't go into government. |
| Patriotic Statements | Our country needs a truly great leader, and we need a truly great leader now. |
| Social Issues | I have so many fabulous friends who happen to be gay, but I am a traditionalist. |
| Supreme Court | The ideal Supreme Court Justice would be Scalia reincarnated. |
| Taxes | If you tax something you get less of it. It's as simple as that. The more you tax work, the less people are willing to work. The more you tax investments, the fewer investments you'll get. This isn't rocket science. |
| Trade | We have very unfair trade with China. |

Respondents were informed that the statements displayed on the MaxDiff screens could be said by a political candidate, but were not initially told the statements were real quotes from Donald Trump. Respondents were then asked to choose the statement they most agreed with and the statement they least agreed with. An example of the question format is shown in Figure 2.

**Figure 2. MaxDiff Example Screen**



The following statements could be said by a political candidate.

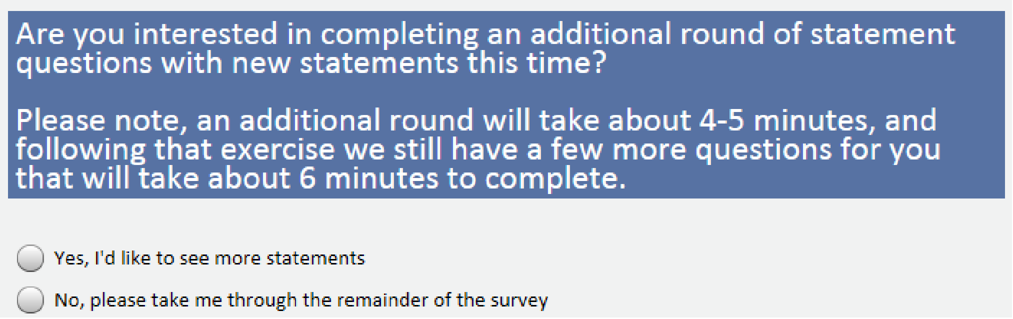
Which of the following statements do you most agree with, and which statement do you least agree with?

(1 of 25)

| Most<br>Agree With | | Least<br>Agree With |
|---|---|---|
| ○ | We've got to bring on the competition. Education reformers call this school choice, charter schools, vouchers, even opportunity scholarships. I call it competition—the American way. | ○ |
| ○ | If you tax something you get less of it. It's as simple as that. The more you tax work, the less people are willing to work. The more you tax investments, the fewer investments you'll get. This isn't rocket science. | ○ |
| ○ | I am a huge fan of the police. I think the police have to be given back power. | ○ |
| ○ | I have so many fabulous friends who happen to be gay, but I am a traditionalist. | ○ |

After completing the first 25 MaxDiff tasks, with exposure to 30 of the 200 items, respondents were asked if they would like to continue the survey by completing an additional round of statement questions with new statements, or if they wanted to simply proceed through the remaining 6 minutes of questions required of all respondents. The question format is displayed in Figure 3.

**Figure 3. Modularization Opt-In Screen**



Are you interested in completing an additional round of statement questions with new statements this time?

Please note, an additional round will take about 4-5 minutes, and following that exercise we still have a few more questions for you that will take about 6 minutes to complete.

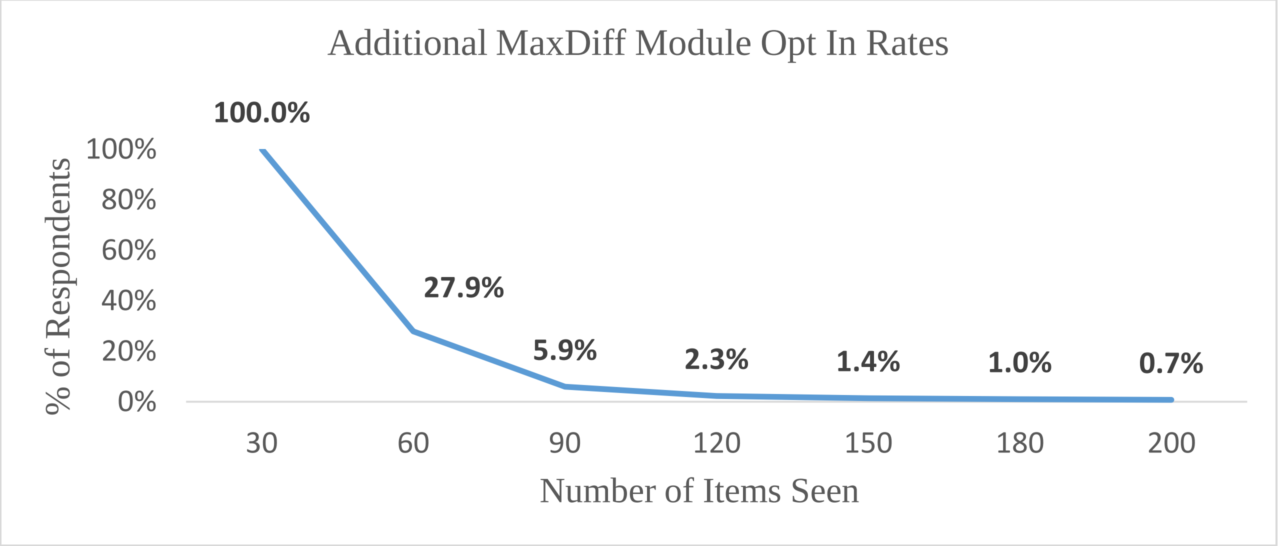
○ Yes, I'd like to see more statements

○ No, please take me through the remainder of the survey

Respondents who chose to opt in to additional rounds of MaxDiff sets were again shown 30 statements (all randomly selected this time) across 25 screens with 4 items per screen. Respondents were able to opt in to up to 6 additional rounds, with the final round consisting of 20 randomly selected statements across 18 screens of 4 statements per screen.

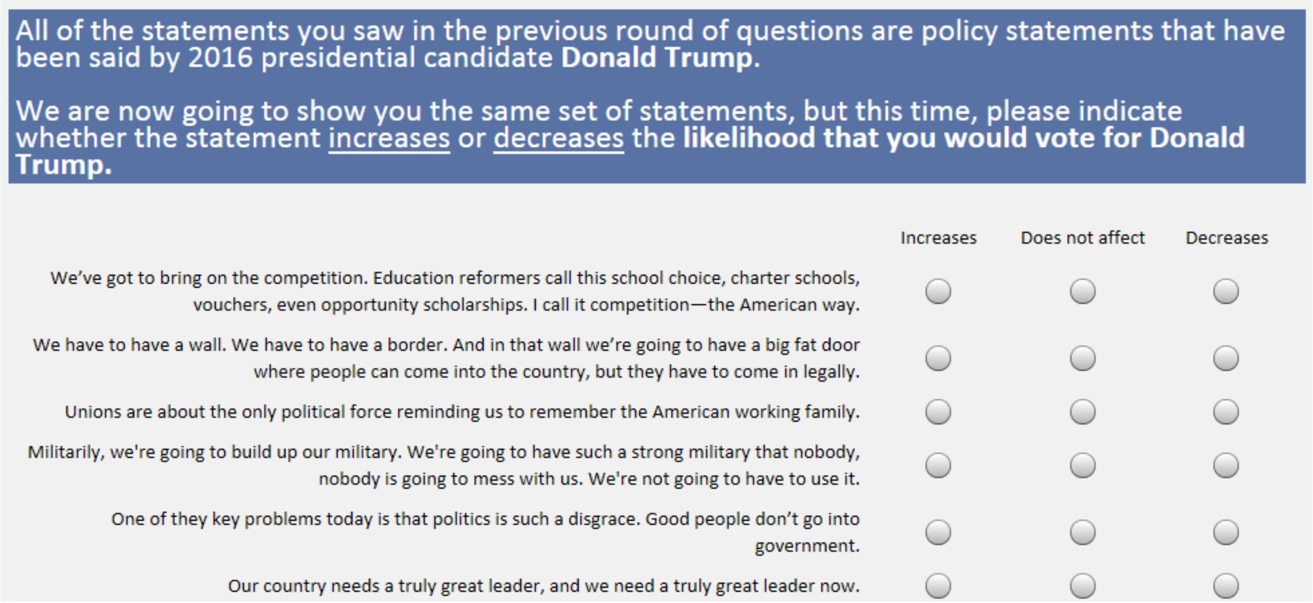
At the first opt in juncture, nearly three quarters of respondents opted out of additional rounds of statements, instead finishing the remainder of the required survey questions and completing the study with only one MaxDiff module. A little over a quarter of respondents proceeded to opt in to an additional round, with the percentage of respondents opting in to additional rounds dropping quickly, as displayed in Figure 4. In the end, a total of only 10 respondents out of n=1,500 were exposed to all 200 statements. This resulted in a massive amount of missing data for our test.

**Figure 4. Opt In Rates**



As soon as a respondent opted out of additional MaxDiff modules (whether it be following the first module or after completing all 7 modules), respondents were exposed to a threshold question. At this point, we informed respondents that all of the statements seen previously had been said by 2016 presidential candidate Donald Trump. We then showed respondents a grid of the 30 statements seen in the first MaxDiff module, and asked them to indicate, for each individual statement, whether it increases, decreases, or does not affect that respondent's likelihood of voting for Donald Trump. These responses were then used to create two anchors using the Direct Binary approach presented by Kevin Lattery (2010). The first anchor distinguished which statements increase respondent likelihood of voting for Trump, while the second anchor distinguished those statements that decrease respondent likelihood of voting for Trump. The exact question wording can be seen below in Figure 5.

**Figure 5. Direct Binary Threshold Question**



All of the statements you saw in the previous round of questions are policy statements that have been said by 2016 presidential candidate **Donald Trump.**

We are now going to show you the same set of statements, but this time, please indicate whether the statement increases or decreases the **likelihood that you would vote for Donald Trump.**

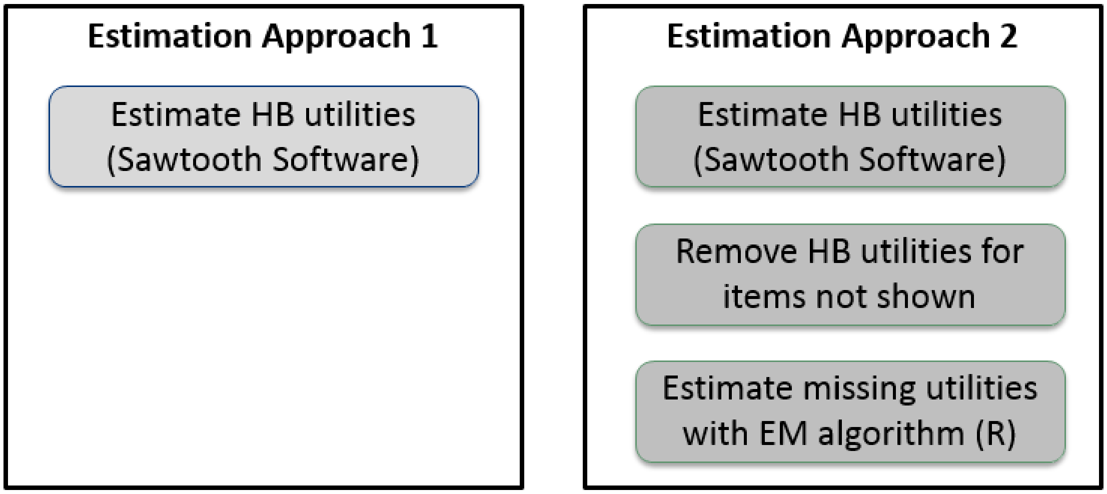
| | Increases | Does not affect | Decreases |
|---|---|---|---|
| We've got to bring on the competition. Education reformers call this school choice, charter schools, vouchers, even opportunity scholarships. I call it competition—the American way. | ○ | ○ | ○ |
| We have to have a wall. We have to have a border. And in that wall we're going to have a big fat door where people can come into the country, but they have to come in legally. | ○ | ○ | ○ |
| Unions are about the only political force reminding us to remember the American working family. | ○ | ○ | ○ |
| Militarily, we're going to build up our military. We're going to have such a strong military that nobody, nobody is going to mess with us. We're not going to have to use it. | ○ | ○ | ○ |
| One of they key problems today is that politics is such a disgrace. Good people don't go into government. | ○ | ○ | ○ |
| Our country needs a truly great leader, and we need a truly great leader now. | ○ | ○ | ○ |

Once the experimental portion of the study was completed, respondents answered a few more questions about their views on the trustworthiness of various forms of media, and then concluded the survey by answering a set of demographic questions.

After collecting all data, we discovered that nearly 75% of respondents were lacking data for 85% of the statements tested. This is where we began our analysis to see whether HB estimation or EM imputation performs better in generating utilities where data are missing.

We started with two separate sets of utilities: one individual level model (anchors included) run in Sawtooth Software using the default HB estimation, and a second model in which we removed the anchored HB utilities generated by Sawtooth Software for items that a given respondent did not see, and then estimated those blanks using the EM algorithm in R (Soft Impute Package).

**Figure 6. Estimation Approaches Used**


Estimation Approach 1
Estimate HB utilities (Sawtooth Software)
Estimation Approach 2
Estimate HB utilities (Sawtooth Software)
Remove HB utilities for items not shown
Estimate missing utilities with EM algorithm (R)

These two sets of utilities, HB only and HB+EM, were used in performing several types of analyses, and the results were compared.

Following some of our initial analysis, we decided to recontact a portion of the original respondents to collect some additional data to serve as a baseline in comparing the utilities estimated using HB and those replaced with imputed EM utilities.

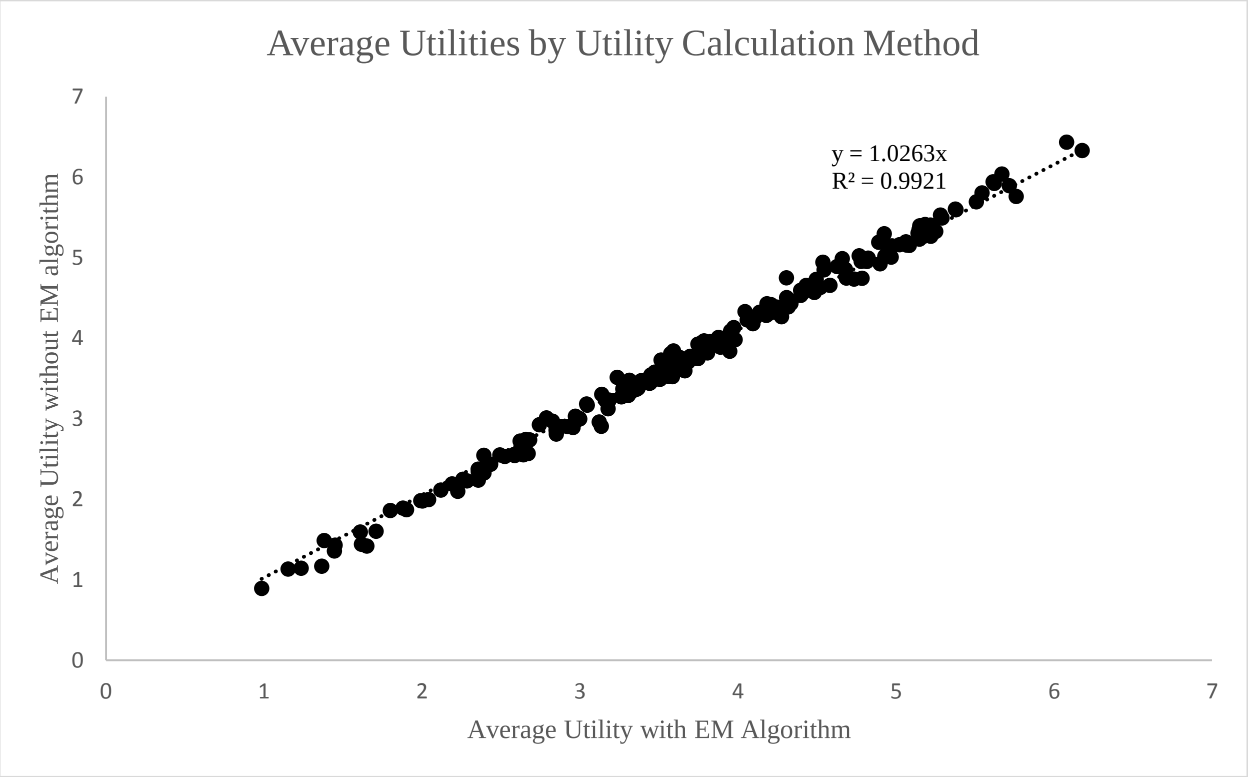
In early July we recontacted 310 respondents for a total of n=200 completes. We used the exact same questionnaire, simply removing those candidates who had at that point dropped out of the race. Respondents only saw one MaxDiff exercise with no modularization. Like the first study, the MaxDiff included 30 statements and consisted of 25 screens with 4 items displayed per task. The design contained 30 statements that the recontacted respondents had not been exposed to in the original survey.

Collecting this data gave us a baseline of "truth" with which to compare the HB only and HB+EM utilities that had previously been estimated with zero direct information about the 30 statements in question.

## RESEARCH RESULTS

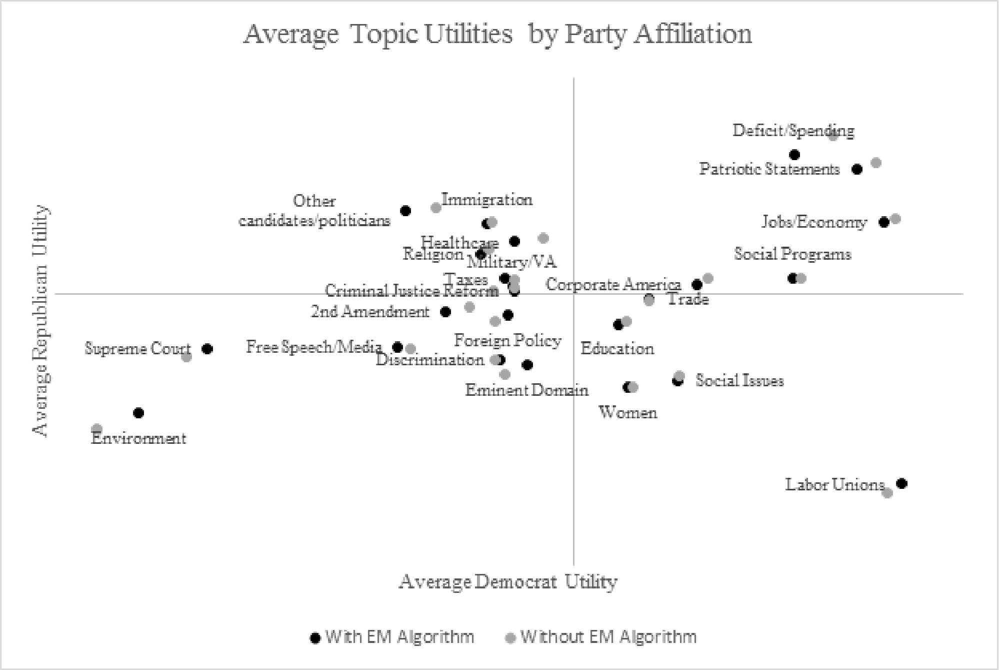
When the two sets of utilities (with and without the EM algorithm) are compared at the aggregate level we end up with very similar results. Figure 7 below plots the mean of each of the 200 items with and without the EM algorithm to estimate the utilities of the missing items. The correlation is very close to 1 (r=.99), which indicates that at the aggregate level using the EM algorithm to estimate the individual level utility scores isn't meaningfully different from the HB estimated utilities.

**Figure 7. Aggregate Utility Comparison**



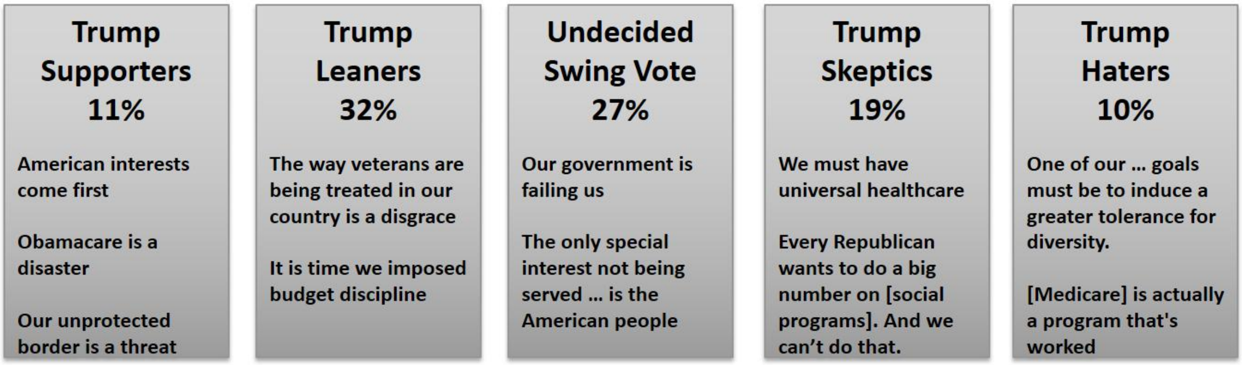
This result holds up even when looking at the key segments of the data such as party affiliation even when not using these variables as covariates. The aggregate utilities also make intuitive sense within these segments of the population. Figure 8 shows the aggregate utilities both with and without the EM algorithm by party affiliation, aggregating the items by topic for an easier read. In particular, you can see that Republicans do not like Trump's statements on labor unions, while the Democrats do. To a lesser extent the same is true with social issues where Trump is considered more moderate than many Republicans. Both parties like Trump's statements on the economy, jobs, and general patriotic statements. The Republicans like his statements on the opposition (in particular Hillary Clinton), immigration, and taxes more than the Democrats. Neither party is particularly enamored with his environmental stances or the delay in the Supreme Court nomination (although Republicans prefer that delay to the Democrats). These trends hold regardless of if the EM algorithm is added to the normal HB scores. However, we do notice more volatility in the segmentation analysis than when we look at the two in aggregate.

**Figure 8. Topical Aggregate Utility Comparison by Political Affiliation**



While the aggregate results are nice, the main reason to do these individual level models is to operate at the individual level rather than the aggregate level. In particular, many times these models are used for segmentation purposes. We ran both sets of utilities with and without the EM algorithm through a Latent Class Segmentation (using Latent Gold) and as a standard k-means cluster analysis. While we looked at 3–7 segment solutions in both cases, we preferred the five segment solution to tell the story of attitudes towards Trump. Figure 9 shows the results of the segmentation. Surprisingly, these results were similar regardless of whether or not the anchors were included in the utility calculations.

**Figure 9. Top Statements by Each Segment**



**Trump Supporters 11%**

American interests come first

Obamacare is a disaster

Our unprotected border is a threat

**Trump Leaners 32%**

The way veterans are being treated in our country is a disgrace

It is time we imposed budget discipline

**Undecided Swing Vote 27%**

Our government is failing us

The only special interest not being served … is the American people

**Trump Skeptics 19%**

We must have universal healthcare

Every Republican wants to do a big number on [social programs]. And we can’t do that.

**Trump Haters 10%**

One of our … goals must be to induce a greater tolerance for diversity.

[Medicare] is actually a program that's worked

While every segment of respondents had statements that they related with, the overall driver of the segment assignment was the anchor in the utilities (increased/decreased likelihood of voting for Trump). Those who supported Trump indexed high on his statements for America’s interests coming first in foreign policy, his statements against the current Obama administration, and strong immigration policy. Those who opposed Trump agreed with more of his statements on the importance of tolerance towards minorities and how current social programs work. While they might agree with these statements he has said, they probably do not agree that Trump exemplifies these statements, as information that Trump was the author of the statement did not cause these respondents to be more likely to vote for him. In fact, many respondents for whom some of the statements resonated simultaneously indicated that knowing Trump said the statements in fact made them less likely to vote for Trump. It is also interesting to note that the undecided swing voters were most likely to be disgusted with the current government and not feel their interests are being represented. Trump's outsider status gives him a leg up in this group. Full results of the 200 statements by segment can be provided at request to the authors. This story does not change depending on which segmentation is used or if the EM algorithm is included or excluded.

Both segmentation methods and utility calculation methods yielded similar results. When using anchored utilities, the classification into the same segment was between 92% and 96%. Removing the anchor does provide more variation in the segmentation algorithms, but the classification into same segments is still extremely high with the lowest being 80%. Table 4 compares these segmentation results and demonstrates the similarity.

**Table 4. Segment Assignment Similarity by Utility Algorithm and Segmentation Method**

| | Anchored Utilities | | | |
|---|---|---|---|---|
| | EM | No EM | EM | No EM |
| | K-Means | K-Means | Latent Class | Latent Class |
| EM K-Means | 100% | 96% | 94% | 92% |
| No EM K-Means | 96% | 100% | 94% | 94% |
| EM Latent Class | 94% | 94% | 100% | 96% |
| No EM Latent Class | 92% | 94% | 96% | 100% |
| | Unanchored Utilities | | | |
| | EM | No EM | EM | No EM |
| | K-Means | K-Means | Latent Class | Latent Class |
| EM K-Means | 100% | 80% | 85% | 83% |
| No EM K-Means | 80% | 100% | 82% | 94% |
| EM Latent Class | 85% | 82% | 100% | 87% |
| No EM Latent Class | 83% | 94% | 87% | 100% |

The final test we did used the recontact data to assess the ability of the two sets of utilities to predict future tasks by the same respondents. We used the individual utilities from each model to see which was better at predicting how respondents would react to these tasks where they had not seen any of the items using hit rates. We compare both of these models to an aggregate logit model to see what, if any, predictive lift we get at the individual level. Table 5 confirms that adding in the EM algorithm on top of the HB algorithm does not boost predictive validity. However, both methods do slightly better than the aggregate logit model, and all models do significantly better than just random chance with no information.

**Table 5. Predictive Comparison Using Hit Rates by Model Type**

| | Best Tasks | Worst Tasks |
|---|---|---|
| EM Anchored | 42.9% | 44.1% |
| EM Unanchored | 42.6% | 43.4% |
| No EM Anchored | 42.3% | 44.0% |
| No EM Unanchored | 42.7% | 43.4% |
| Aggregate Logit | 38.7% | 40.2% |

## CONCLUSION

It is clear that respondents get fatigued when completing a large series of MaxDiff tasks. Most of our respondents decided to leave after 25 MaxDiff tasks rather than continue evaluating more political statements. However, not everyone was equally fatigued. Giving respondents the option to complete additional MaxDiff tasks did give us more complete data on a few respondents who likely had higher interest in the topic. Finding a way to estimate unseen items is essential to keeping an engaged audience when confronted with a large number of items to evaluate.

Trying to predict individual level utilities where a respondent has not seen the item in a MaxDiff exercise is uncertain at best. While the HB algorithm will estimate individual level utilities, there is a lot of shrinkage going on to the aggregate model. While we attempted to add in covariates to help guide the shrinkage, the amount of time it took to run the model with these covariates was prohibitive. Removing the utilities for items not seen in the tasks and imputing them with an EM algorithm yields very similar results. The two algorithms produce similar results not only at the aggregate level, but on the individual level when looking at segmentation composition. Both types of utilities give a slight boost in hit rates for items when a respondent sees new items in a recontact when compared to the aggregate logit model.

One important advantage of HB is that it has an estimate of variance in the draws, whereas the EM algorithm only provides point estimates. However, industry market simulators commonly only use the point estimates, which do not allow for the full advantage of HB. While calculating individual level utilities might not boost predictive performance much above the aggregate model, it can still be useful in these simulators that use the point estimates. In particular, individual level utilities allow simulators to recalculate preference for subgroups on the fly without having to calculate new utilities. The aggregate logit model would need to be recalculated for every desired segment of the population rather than just taking a new average of existing individual utilities. The extra step of removing the HB utilities for unseen items and replacing with an EM algorithm doesn't seem to be worth the effort. We would not recommend it in the future.

There are many avenues researchers could take to build on this work. In particular, with political surveys there are known covariates (such as party affiliation and past voting behavior) that could likely inform the models on respondent attitudes towards statements that were not seen. Using a Sparse MaxDiff instead of an Express MaxDiff to get some information on most of the statements instead of a lot of information on only some of the statements could yield much better results. Lastly, even using the results of a segmentation algorithm (either the ones used or a cluster ensemble algorithm) might improve the model's hit rates on the individual level by adding it in as a covariate.

We would also prefer to do the recontact closer to the initial survey as several big news items occurred between the two waves of data collection. In particular, presidential candidates dropped out of the race, which could have changed respondent's opinions on the statements and of Trump in general. This delay could have contributed to the overall lower hit rates of all the models when predicting respondent choices in the recontact data.

Kelsey White

Paul Johnson

## Works Cited

Fairchild, K., Orme, B. & Schwartz E. (2015). Bandit Adaptive MaxDiff Designs for Huge Number of Items. *2015 Sawtooth Software Conference Proceedings.* pp. 105–117.

Hendrix, P., & Drucker D. (2007). Alternative Approaches to MaxDiff with Large Set of Disparate Items—Augmented and Tailored MaxDiff. *2007 Sawtooth Software Conference Proceedings.* pp. 169–188.

Lattery, K. (2007). EM CBC: A New Framework for Deriving Individual Conjoint Utility by Estimating Responses to Unobserved Tasks via Expectation-Maximization. *2007 Sawtooth Software Conference Proceedings.* pp. 127–138.

Lattery, K. (2010). Anchoring Maximum Difference Scaling Against a Threshold—Dual Response and Direct Binary Methods. *2010 Sawtooth Software Conference Proceedings.* pp. 77–90.

Orme, B. (2005). Accuracy of HB Estimation in MaxDiff Experiments. *Sawtooth Software Research Paper*, available at www.sawtoothsoftware.com/download/techpap/maxdacc.pdf.

Orme, B. (2006). Adaptive Maximum Difference Scaling, *Sawtooth Software Research Paper*, available at www.sawtoothsoftware.com/support/technical-papers.

Wirth, R. & Wolfrath, A. (2009). Using MaxDiff for Evaluating Very Large Sets of Items. *2009 Sawtooth Software Conference Proceedings.* pp. 59–78.

# THE RESEARCHER'S PARADOX: A FURTHER LOOK AT THE IMPACT OF LARGE-SCALE CHOICE EXERCISES

***MIKE SERPETTI***
***CLAIRE GILBERT***
*GONGOS*
***MEGAN PEITZ***
*SAWTOOTH SOFTWARE*

## ABSTRACT

MaxDiff is ideal for determining preference order, yet becomes problematic for large item-sets. While sparse and express methods address these issues, not enough large item-set research exists. This research explores and validates both methods, with real respondents, on a set of 100 items to determine which is better.

## BACKGROUND

Clients continue to push the envelope when it comes to increasing the length of item lists in a MaxDiff (Maximum Difference Scaling or Best-Worst Scaling) study. For traditional MaxDiff, Sawtooth Software has suggested that each item should be seen at least three times per respondent for accurate, individual-level estimation. This rule of thumb means longer lists of items (i.e., >50) equate to more survey screens, which can lead to respondent fatigue and bad data. It is our duty as researchers to consider how valuable our respondents' time is. While researchers want more data to ensure better estimation, they have to balance this desire with what respondents are willing to provide.

MaxDiff is a superior technique for determining preference order among a list of items. This methodology not only ranks a list of items, but also reveals the magnitude of difference between ranked items. Additionally, MaxDiff has been valuable in other research applications such as TURF or Segmentation, and has proven its use in previous research for up to 30 to 40 items.

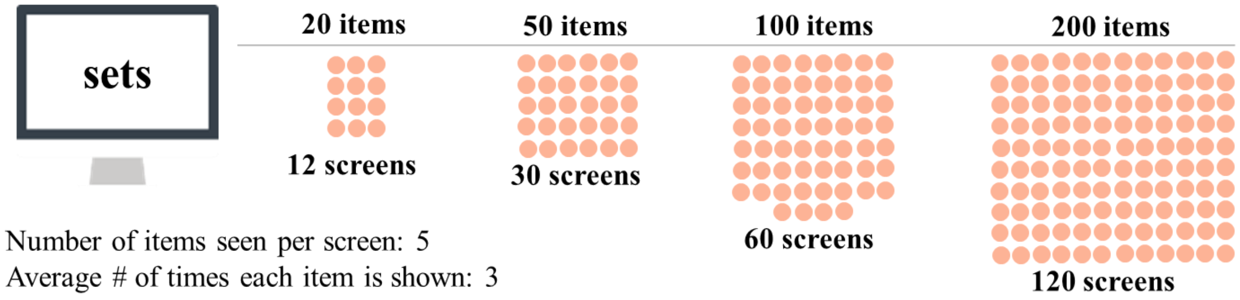
Under the traditional MaxDiff method, the more items included in the exercise, the longer the survey will be. The standard formula for determining the number of sets is shown below in Equation 1. In this formula, the goal is for each item to be seen at least three times, on average, by each respondent.

**Equation 1. The Numbers of MaxDiff Screens Required for Analysis**

$$\text{Number of sets} = \frac{\text{Number of items being tested} * 3}{\text{Number of items shown per set}}$$

The number of screens increases quickly as the number of items being tested increases. Assuming the formula above, along with showing each respondent five items per set, Figure 1 indicates how long and taxing this exercise can be for a respondent.

**Figure 1. The Number of Set Required to Test per Number of Items**



One of the major issues facing a long MaxDiff exercise is respondent fatigue, resulting in bad data that can bias final results. **This paper encourages researchers to address this issue and explores if there is a way to obtain accurate estimation without taxing respondents in large-scale choice exercises.**

## Previous Research

Over the past decade, a number of practitioners have investigated different ways of analyzing large-scale MaxDiff exercises. These include:

- Orme—Adaptive MaxDiff
- Hendrix/Drucker—Augmented MaxDiff, Tailored MaxDiff
- Wirth/Wolfrath—Express MaxDiff, Sparse MaxDiff
- Orme/Fairchild/Schwartz—Bandit Adaptive MaxDiff

Most of the research presented in this paper stems from the work presented by Wirth and Wolfrath in their exploration of sparse and express MaxDiff. The theory behind sparse MaxDiff is to show respondents each item less often than they would see in the traditional method. For example, under the traditional method, respondents are usually shown each item an average of three times; whereas with the sparse methodology, respondents might only see each item fewer times on average.

On the other hand, the express methodology uses a different approach by showing respondents a subset of the items being tested. For example, a study with a list of 100 items is broken up into smaller subsets (e.g., 30) where each respondent receives one subset. Then, Hierarchical Bayes is used to stitch the data together into a set of utilities that can be used for analysis. The main advantage of both of these methodologies is their ability to test a large number of items without burdening respondents with a longer survey.

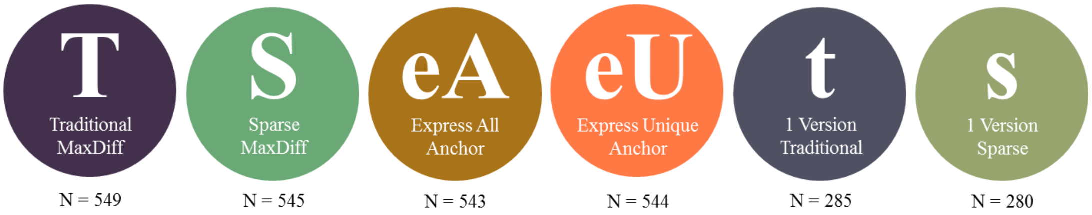
Although there has been a lot of research in the past decade, many of these studies have either been simulations or have contained lists of items that were not excessive in nature (e.g., Wirth and Wolfrath tested 60 items). The purpose of this paper is to determine if an excessive number of items (i.e., 100 items) can be tested under three different MaxDiff methods: traditional MaxDiff, sparse MaxDiff, and express MaxDiff.

## Research Design

A MaxDiff experiment containing 100 items was fielded among 2,746 respondents through GMI's Lightspeed panel. The survey tested the importance of a wide array of items related to ice cream. It contained items sure to resonate with many respondents, such as: "it tastes good," "it has good add-ins," and "it's creamy." However, it also tested the importance of more obscure items, including: "the packaging is biodegradable," "the packing is see-through," and "the ice cream is made with stevia."

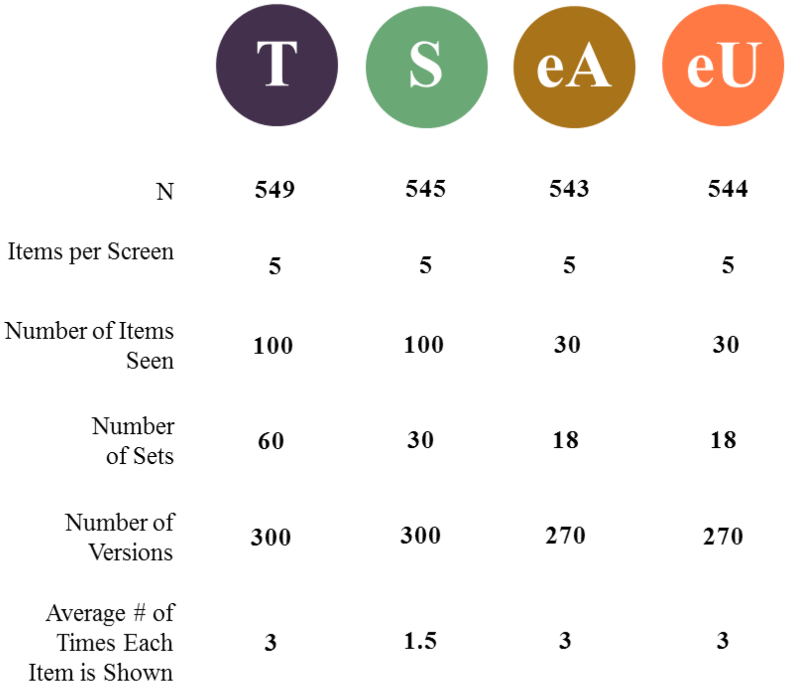
Respondents were screened to be between the ages of 18 and 65, to have at least half of the responsibility for grocery shopping in their household, and have had purchased or consumed packaged ice cream within the past six months. Each respondent was then randomly assigned to one of six MaxDiff methods, shown below in Figure 2. When data collection was complete, the data was weighted to the same demographic proportions across all 6 cells.

**Figure 2. MaxDiff Methods Tested**



Each of the MaxDiff methods shown in Figure 2 was set up to be unique and depended on the overall number of items seen, the number of sets and versions shown and the average number of times each item was shown per respondent. A more detailed description of the first four methods appears in the figure below (Figure 3).

**Figure 3. MaxDiff Method Details**



| | T | S | eA | eU |
|---|---|---|---|---|
| N | 549 | 545 | 543 | 544 |
| Items per Screen | 5 | 5 | 5 | 5 |
| Number of Items Seen | 100 | 100 | 30 | 30 |
| Number of Sets | 60 | 30 | 18 | 18 |
| Number of Versions | 300 | 300 | 270 | 270 |
| Average # of Times Each Item is Shown | 3 | 1.5 | 3 | 3 |

As previously referenced, two other MaxDiff versions were tested. The first was a 1-version traditional MaxDiff exercise (n = 285) and the other a 1-version sparse MaxDiff exercise (n = 280). Both of these methods were randomly assigned to respondents and were used as an out-of-sample validation test against the main, three hundred versions of both the traditional and sparse MaxDiff methods. Another note of importance is that both of the express MaxDiff methods contained the same exact design.

In addition to testing the accuracy of estimation for each MaxDiff method, another goal of this research was to test if applying Kevin Lattery's Direct Anchoring Method could improve accuracy measures across these different methods. Respondents who saw the traditional and sparse were asked a follow-up Direct Anchoring question on all items. This was a multi-punch question that also offered a "none of the above" option. This same process was done for the express all anchor respondents, even if their specific MaxDiff did not include these items. On the other hand, express unique anchor respondents were only shown items that they rated in their actual MaxDiff exercise.

## Accuracy Analysis

A hit rate was calculated using a holdout task to determine the accuracy of each method in identifying preference at the individual level. The hit rate is the proportion of the holdout ranks that match the utility ranks across the entire population. In the holdout task, respondents were asked to rank five statements from the MaxDiff in order of preference. They did this for five unique batteries—resulting in 5 holdout tasks, or more specifically, 25 statements that could be compared to the rank order of the utilities. For comparison, the utilities for the 25 statements were ranked in order within each of the 5 groups. Each person had 25 pairs of ranks that could either match or not.

Due to the large item list, some utilities were very close to one another. To accommodate these very close scores that could essentially be ties in the respondent's mind, a confidence interval was developed to apply to the hit rate scores. This process allows very close utilities, that overlap within their margin of error, to flip ranks and be considered a successful match. The 95% confidence interval for each of the 25 statements was created using the standard deviation of all utilities for the statement. Bounds of 95% were applied to the utility to create a range of utilities (1.96*sd). These ranges were ranked; those that overlapped could count as a success in either rank, while those that had no overlap were assigned their rank.

Using either hit rate calculation, traditional MaxDiff is the most accurate. Sparse follows as a close second, while both express methods are less accurate, seen below in Figure 4.

**Figure 4. Pure and Confidence Interval Hit Rates**

|  | T | S | eA | eU |
|---|---|---|---|---|
| Pure Hit Rate | 38% | 34% | 32% | 30% |
| Confidence Interval Hit Rate | 58% | 52% | 49% | 47% |

## Correlation Analysis

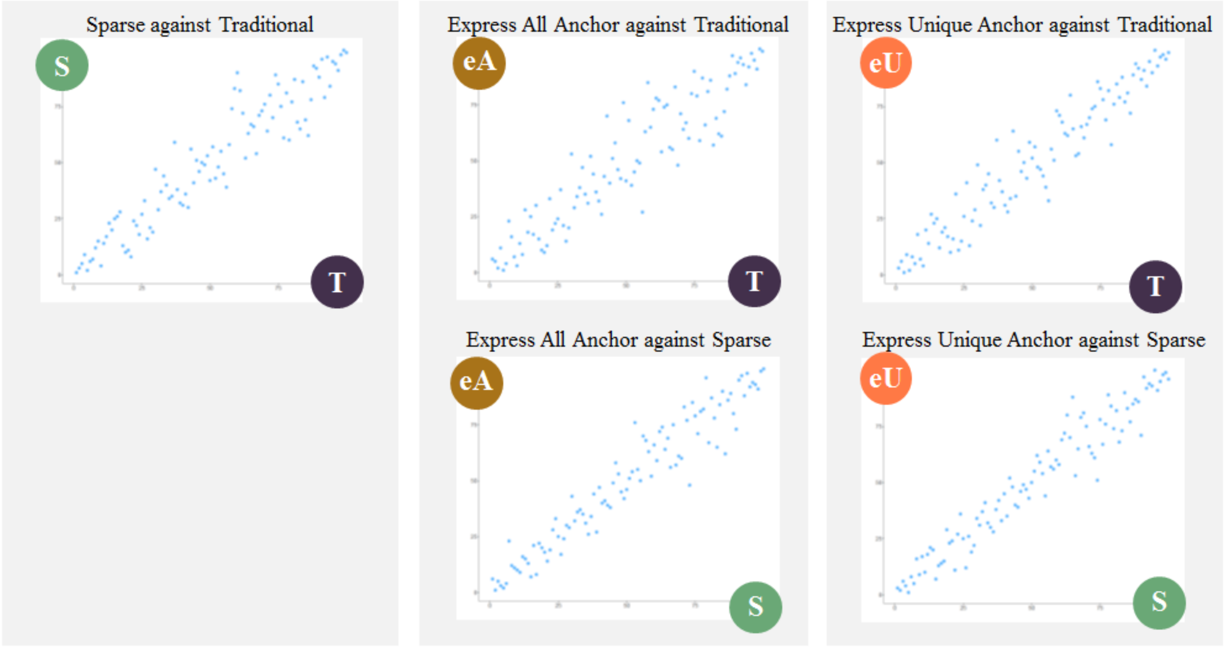
Another measure of accuracy is the correlation between the utility ranks and the ranks from the holdout tasks at the individual level. This measures the relationship between the ranks of the holdouts and the ranks of the generated utilities. Using Spearman's rank correlation, a metric was calculated for each of the five holdout tasks, comparing the five holdout ranks to their corresponding utility's ranks. This results in five correlations for each respondent. The average correlation was taken to create one measure for each respondent. Similar to results from hit rate calculations, the traditional method shows the strongest relationship between the holdout and MaxDiff exercise with sparse falling just behind, followed by both express methods shown in Figure 5.

**Figure 5. Spearman's Rank Correlations for Holdout, Traditional, and Sparse Utility Ranks**

| | T | S | eA | eU |
|---|---|---|---|---|
| Correlation to Holdout Ranks | .49 | .41 | .35 | .34 |
| Correlation to Traditional Ranks | | .95 | .93 | .96 |
| Correlation to Sparse Ranks | .95 | | .96 | .97 |

When items ranks are plotted, sparse ranks align better to the traditional ranks than the express, however sparse and express align to one another better than either does to the traditional, shown below in Figure 6.

**Figure 6. Utility Ranks Plotted against Different Designs**



Sparse against Traditional

Express All Anchor against Traditional

Express Unique Anchor against Traditional

Express All Anchor against Sparse

Express Unique Anchor against Sparse

## Out-of-Sample Validation

Another way to test the models' performance is out-of-sample validation. Two sample groups were withheld from the estimation process, each consisting of n=250$^+$ respondents. One group saw a 1-version traditional MaxDiff, resulting in 60 fixed sets for this out-of-sample data. The other group saw a 1-version sparse MaxDiff, resulting in 30 fixed sets of holdout data. Design details are shown below in Figure 7.

**Figure 7. Details for the 1-Version Out-of-Sample Validation Versions**

| | t | s |
|---|---|---|
| N | 285 | 280 |
| Items per Screen | 5 | 5 |
| Number of Items Seen | 100 | 100 |
| Number of Sets | 60 | 30 |
| Number of Versions | 1 | 1 |
| Average # of Times Each Item is Shown | 3 | 1.5 |

Using model estimates from each of the four methods (T, S, eA, eU), share of preference (SOP) simulations were compared to the counts of the "best" items in the holdout exercises. It should be mentioned that the scale factor for each of the four models was also tuned as to minimize the mean absolute error (MAE) within each model.

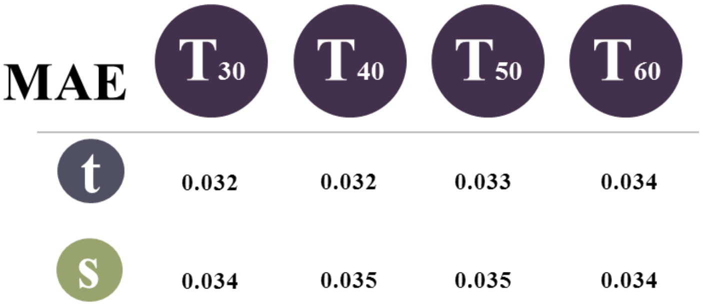
The table below shows the MAE scores for each of the four methods at the aggregate level; the goal is to minimize the MAE (*Figure 8*) with a typical MAE found between .02 and .04. Although the differences in MAE are likely not significant, they do suggest that sparse is a viable alternative to the traditional.

**Figure 8. Mean Absolute Error Scores for Each Method Using the 1-Version Out-of-Sample Validation**

| MAE | T | S | eA | eU |
|---|---|---|---|---|
| t | 0.034 | 0.030 | 0.038 | 0.036 |
| s | 0.034 | 0.030 | 0.038 | 0.039 |

Another hypothesis was that increasing the number of tasks used in model estimation would decrease the MAE (i.e., more sets result in less error). If this isn't true, one might assume the quality of an individual's response degrades over time. Therefore, using data from the traditional only, four separate utility estimates were created using the first 30 tasks; then 40; then 50; then all 60. The table suggests that even as the number of tasks used in the utility estimation increases, there is not a significant decrease in MAE (Figure 9). This is another indication that sparse or express methods may be just as, if not more, effective than traditional in utility estimation for large item sets.

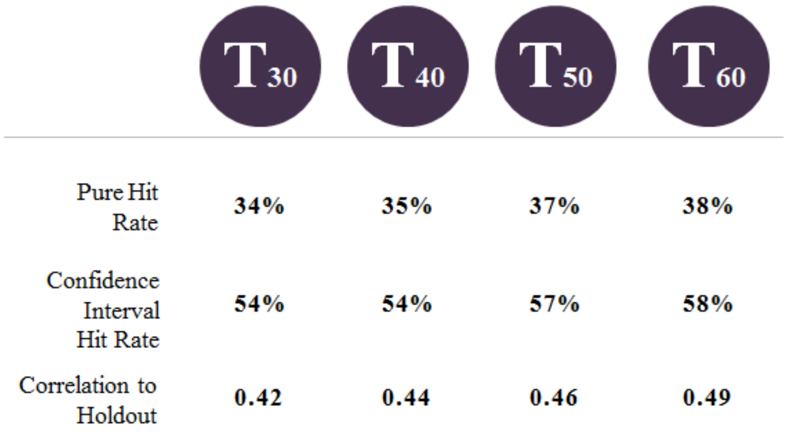
**Figure 9. Mean Absolute Error Scores for Traditional Estimating Using the First 30, 40, 50, and 60 Tasks and the 1-Version Out-of-Sample**



| MAE | T30 | T40 | T50 | T60 |
|---|---|---|---|---|
| t | 0.032 | 0.032 | 0.033 | 0.034 |
| s | 0.034 | 0.035 | 0.035 | 0.034 |

The traditional method is more accurate than the proposed alternatives, however, it is a much longer exercise to include in a survey. In order to understand the increase in accuracy relative to the number of tradeoff sets evaluated by respondents, hit rates and correlations were calculated for the first 30, 40, and 50 sets seen. As more tasks are estimated using Hierarchical Bayes, only slight increases among in-sample hit rates occur. This further suggests that the gain in adding more sets may not offset the loss through respondent fatigue and satisfaction. Hit rates and correlations suggest that sparse is approximately as accurate as a traditional method with a limited number of sets (Figure 10 and Figure 4).

**Figure 10. Hit Rates and Correlation to Holdout for Traditional Using Only First 30, 40, 50 and All 60 Sets**



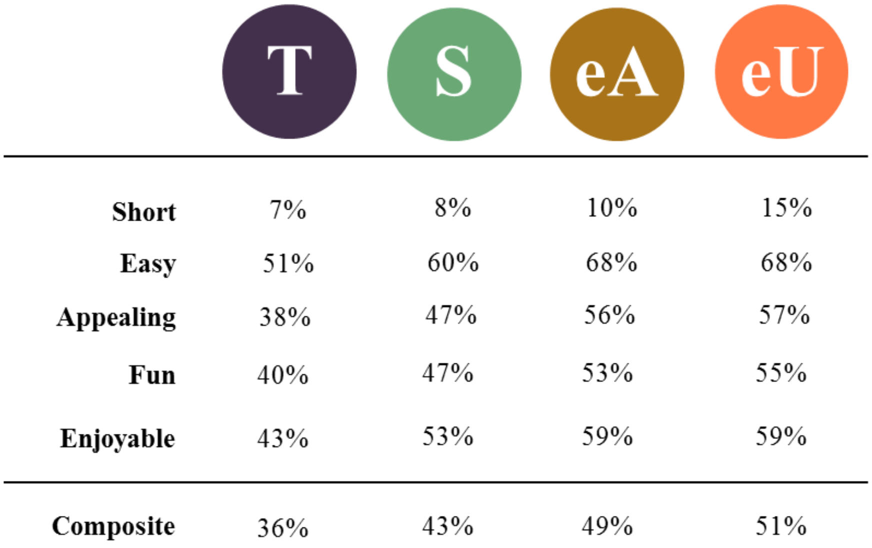
| | T30 | T40 | T50 | T60 |
|---|---|---|---|---|
| Pure Hit Rate | 34% | 35% | 37% | 38% |
| Confidence Interval Hit Rate | 54% | 54% | 57% | 58% |
| Correlation to Holdout | 0.42 | 0.44 | 0.46 | 0.49 |

## Respondent Satisfaction

Although the scientific purpose of this paper is to focus on the statistical accuracy of each of the MaxDiff methods tested, the respondent experience is a very important piece of the puzzle that cannot be overlooked. Respondents who are disengaged throughout the course of the survey or who become dissatisfied with the length of the survey can have a substantial impact on the overall results. Disengaged respondents can lead to responses with a large amount of directional noise—without a researcher being able to discover the issue—or the need to clean a large number of respondents out of a data set.

In order to gauge satisfaction, respondents were asked a series of seven-point semantic differential questions about the MaxDiff exercise they took. Some examples of these questions included: was the MaxDiff portion short or long, was it fun or dull, and was the MaxDiff enjoyable or not. The results with regard to respondent satisfaction are shown below in Figure 11 and indicate a top-two box score with regard to the semantic differential question (Figure 11).

**Figure 11. Top Two Box Scores for Respondent Satisfaction on Each Attribute**

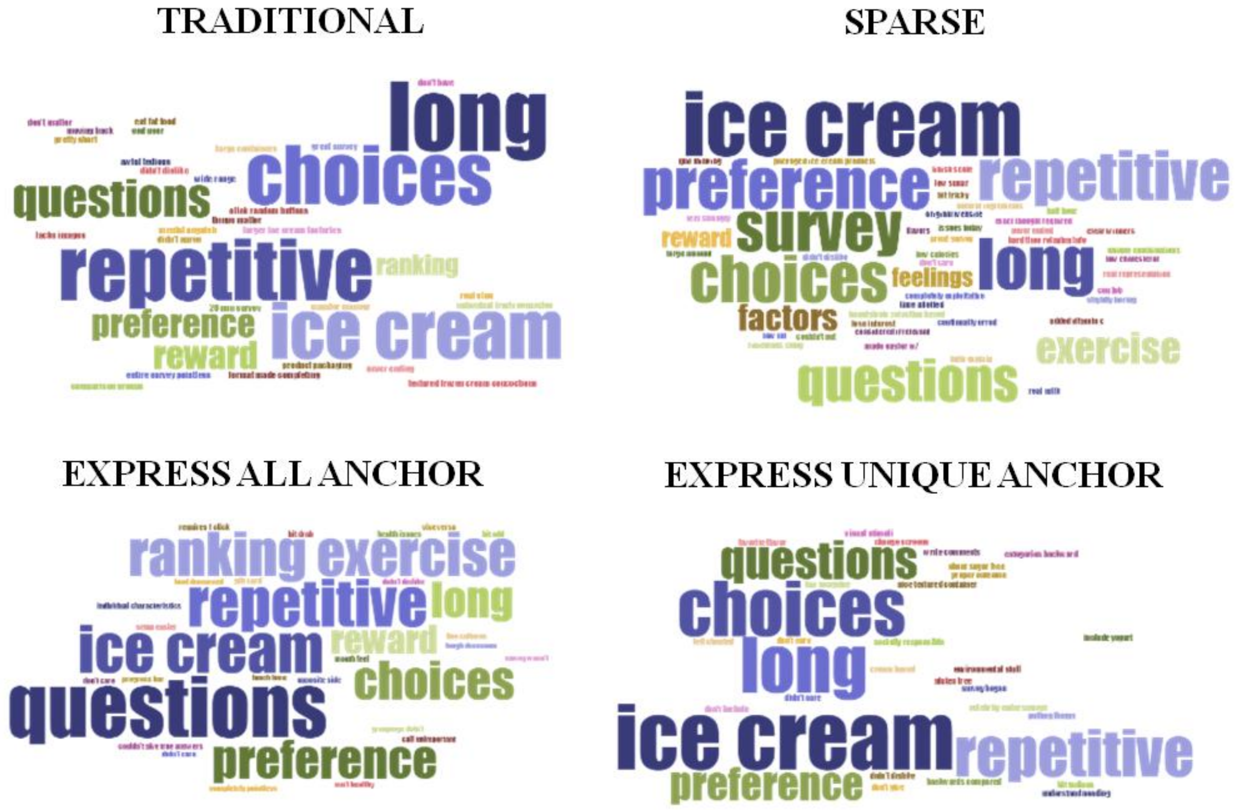


| | T | S | eA | eU |
|---|---|---|---|---|
| **Short** | 7% | 8% | 10% | 15% |
| **Easy** | 51% | 60% | 68% | 68% |
| **Appealing** | 38% | 47% | 56% | 57% |
| **Fun** | 40% | 47% | 53% | 55% |
| **Enjoyable** | 43% | 53% | 59% | 59% |
| **Composite** | 36% | 43% | 49% | 51% |

These results show a positive trend from those who participated in the longer, traditional MaxDiff to those who were shown the shorter, express MaxDiff. Respondent satisfaction was lower for those who saw the traditional; they explained that the survey was long, hard, less appealing, and duller. These results improved for those respondents who were assigned the sparse MaxDiff and were the highest with the express. In addition to being shorter, the more satisfied respondents also found the survey easier, more enjoyable, more appealing, and more fun.

Respondents were asked open-end questions about what they liked and disliked in the MaxDiff exercises. These were analyzed using Rapid Automatic Keyword Extraction text analysis. Common themes in the “Like” category included “ice cream” and “choices,” with little variability across MaxDiff methods. However, the “Dislike” category showed more variation between methods, as can be seen in the word clouds in Figure 12. The traditional emphasized the “long” and “repetitive” nature of the exercise, whereas the sparse highlighted “preference” and “choices.” Both of the express methods emphasized the “repetitive” nature of the exercise and the choices available—which makes sense, as they only saw a small subset of the items available. The key takeaway from the text analysis was the traditional method was viewed as “long.”

**Figure 12. Word Clouds for the "Dislike" Open End for Each Method**



TRADITIONAL

SPARSE

EXPRESS ALL ANCHOR

EXPRESS UNIQUE ANCHOR

## "Bad" Respondent Analysis

Another objective of this research was to determine the amount of "bad data" respondents each MaxDiff method yielded. A respondent was determined to be "bad data" if they failed two or more data quality checks or if that respondent admitted to cheating during their survey. The different quality checks implemented throughout the survey included: respondent completion time for their respective MaxDiff, a poor Root Likelihood Score, straight lining on other questions asked throughout the survey, or incorrectly answering a question for which they were told to mark a specific answer. When analyzing the amount of bad data across each of the different designs, the traditional MaxDiff and the express all anchor had the highest percentage removed (16%). The sparse and express unique anchor showed fewer respondents removed, 8% and 12% respectively.

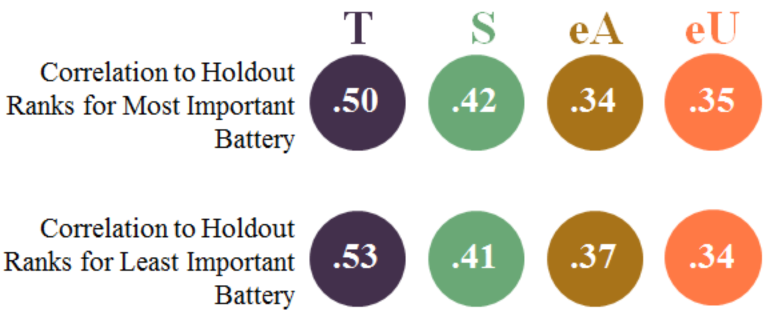
The traditional MaxDiff took respondents a median completion time of 16.4 minutes, while the shorter exercises, sparse and express, took respondents a median time of 9.7 minutes and 5.5 minutes, respectively. This measure was calculated using all respondents who completed the survey, whether they were flagged as a "bad" respondent or not. When focusing on this overall sample (good and bad respondents), the longer, traditional MaxDiff had more respondents considered "speeders" (14%). A speeder is defined as anyone who completed in half the median completion time for their assigned MaxDiff design. On the other hand, the sparse and express MaxDiff respondents had a lower number of speeders (9% for sparse and 10%–11% for the two express methods).

As mentioned earlier, after respondents had finished completing the MaxDiff exercise, they were asked if they thought about cheating. If answered “yes,” they were asked if they actually cheated. Respondents were made aware that they would not be penalized for their truthful answer, would still receive the incentive promised to them, and that they would not face any repercussions from the panel company. With regard to the sparse “bad” respondents, nearly 100% of the data removed was due to respondents admitting to cheating, while the two express methods averaged over 75%. However, when looking at the traditional, under 70% of the bad data was accounted for by cheating, which means that these respondents were more likely to be removed for other reasons (i.e., poor timing, straight lining, etc.). Amongst the data that remained as good data, almost 25% of the respondents in the traditional method considered cheating or straight lining in order to move through the survey more quickly. As the survey length shortened, the percentage of those who considered cheating decreased (sparse: -18%; express: -14%).

## Estimating Importance Values

Previous research by Wirth and Wolfrath states that sparse and express MaxDiff methods perform similarly when estimating the importance of highly ranked items. To evaluate how the various methods performed against the more and less important items, the two holdout ranking tasks that contained the five most important and the five least important items were isolated. Most and least important items were determined from a previous study containing a subset of this item list. These holdout ranks were correlated to the corresponding items’ ranks in order to establish a measure of comparison. This research found that sparse did better than express with the more important items and slightly better than express with the least important items (Figure 13).

**Figure 13. Spearman’s Rank Correlation for Most and Least Important Holdout Batteries**



|  | T | S | eA | eU |
|---|---|---|---|---|
| Correlation to Holdout Ranks for Most Important Battery | .50 | .42 | .34 | .35 |
| Correlation to Holdout Ranks for Least Important Battery | .53 | .41 | .37 | .34 |

To further understand how the alternative methods perform against the traditional MaxDiff, the top 25 ranked items from the traditional were correlated to their corresponding ranks in each of the other three methods. The same was done with the bottom 25 traditional ranks. The story is similar to that of the holdout task as well—sparse does better than express with the top ranked items, but conversely, express seems to do better with the items ranked lowest (Figure 14).

**Figure 14. Correlation to Top and Bottom 25 Ranks**

| | S | eA | eU |
|---|---|---|---|
| Correlation to Top 25 Traditional Ranks | .69 | .62 | .66 |
| Correlation to Bottom 25 Traditional Ranks | .80 | .79 | .86 |

## Additional Techniques Explored for Improving Results

### Education

Inspired by the findings of Sawtooth Software's Turbo Choice event, this paper sought to explore if a respondent's experience would improve if they were educated prior to the MaxDiff exercise. The sample was split into two groups, one of which was educated prior to the MaxDiff exercise with the following text "One of the benefits we like to provide consumers is real time feedback to thank them for taking part in this survey. At the end of this exercise, we will reveal your top 5 attributes that are MOST important to you in terms of packaged ice cream." The other group received no information.

After the MaxDiff exercise, Sawtooth Software's MaxDiff on-the-fly calculations were called upon to show the educated respondents their top 5 ranked items. Respondents were then asked, "How well do these statements align with what is most important to you when buying packaged ice cream?"

Among those that were educated, the traditional method does the best at aligning the estimated top 5 ranks with actual top 5 ranks, followed closely by sparse and then express. A slight lift in hit rate predictability is also observed (Figure 15). However, further investigation is necessary to understand why respondents' answers don't align as well with the express methods. The assumption is that respondents are more critical of their on-the-fly results in the block designs because they had a shorter list of items to evaluate and thus remember more about each item's preference.

**Figure 15. On-the-Fly Alignment with Top Attributes Sought and Hit Rates for the Educated Sample**

| | T | S | eA | eU |
|---|---|---|---|---|
| Do on-the-fly rankings align with your responses (top-two box)* | 93% BEST ALIGNS | 90% | 86% | 85% |
| Pure Hit Rate* | 37% ↓ 1% | 35% ↑ 1% | 32% | 29% ↓ 1% |
| Confidence Interval Hit Rate* | 62% ↑ 4% | 56% ↑ 4% | 53% ↑ 4% | 48% ↑ 1% |

*Results amongst those who were educated

↑↓ Indicate an increase/decrease from original holdout scores

More importantly is the lift in key satisfaction measures by the educated group compared to the non-educated group. In almost every case, the exercise is more appealing, more fun, and more enjoyable, regardless of MaxDiff method (Figure 16). These results would suggest that making respondents aware that they will be provided with their top ranked attributes at the end of the exercise should be implemented wherever possible, especially given the ease of using Sawtooth Software's MaxDiff on-the-fly script language.

**Figure 16. Top Two Box Satisfaction Measures for the Educated Sample**

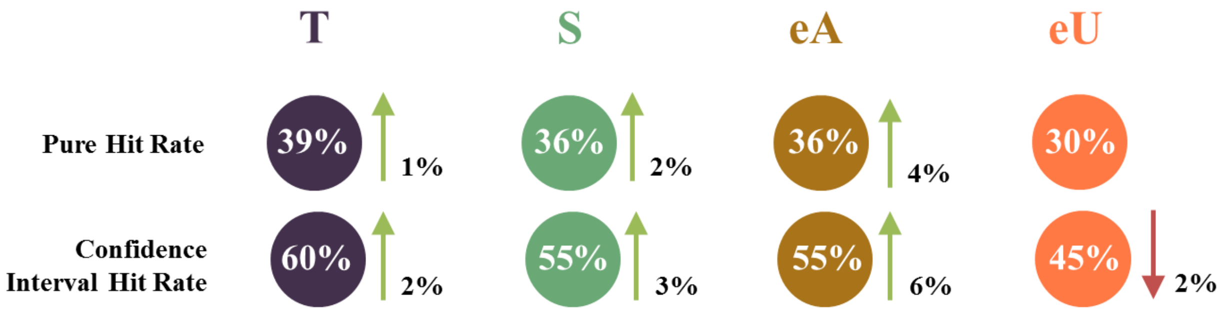
| | | T | S | eA | eU |
|---|---|---|---|---|---|
| Appeal | Educated | 45% | 49% | 59% | 59% |
| | Not Educated | 31% | 44% | 53% | 54% |
| Fun | Educated | 48% | 47% | 55% | 60% |
| | Not Educated | 33% | 47% | 50% | 50% |
| Enjoyable | Educated | 51% | 54% | 65% | 63% |
| | Not Educated | 35% | 52% | 54% | 55% |

## ANCHORING

Because MaxDiff scores are relative and researchers often want to draw a line between the items that are actually important to respondents and those that are not important, this research applied the Direct Binary Anchoring approach developed by Kevin Lattery (Lattery 2011). This method employs the standard MaxDiff questioning followed by a multi-select question at the end that is used for threshold estimation.

Exploring the results shows that including the Direct Binary approach in the analysis has a positive effect on the hit rates of the traditional, sparse, and express methods when all items are used in the anchor question (eA). However, when only the subset of items tested in the express MaxDiff are used in the anchor (eU), there is a slight negative effect (Figure 17). This is likely due to the fact that the anchor in the eU approach doesn't supply any information on items outside of the design. While a random subset of items may work well as the anchor, it is recommended that the subset not be limited to the items tested in the block.

**Figure 17. Hit Rates for All Design Employing Direct Anchoring**



It should also be mentioned that in comparing the position of the anchor, the anchors for the sparse and express all anchor methods were much higher (positions 12 and 13 respectively) than the traditional anchor (position 21). One hypothesis is that with less data, the actual anchor can fluctuate in comparison to the traditional method, which estimates with more information. While it cannot be said which approach is more "correct," there would be a significant difference in the insights drawn from the differing anchor positions depending on the estimation method used.

## Conclusions and Future Research

Overall, looking at large-scale MaxDiff studies, the traditional method is statistically the best when measured by accuracy hit rates amongst holdout tasks. However, this type of study is extremely long and challenging to respondents. It runs the risk of alienating respondents and leads to a significant amount of bad data and wasted sample.

There are alternatives to help solve for the length of traditional MaxDiffs, including either sparse or express. The express method is preferred most by respondents due to its shorter length and more efficient questioning structure. However, the research within this paper demonstrates that it is the statistically least precise of the three methods tested, suffering the most on accuracy. Based on the data gathered during this research, the sparse MaxDiff is able to hold up both ends of the spectrum in terms of data accuracy and respondent satisfaction. Given its strength in both of these areas, sparse MaxDiff is the recommended methodology when dealing with MaxDiff studies that are extremely large in nature.

Further exploration is required to answer the question: if the sparse method in this study only showed each item one time—on average, per respondent—and if the connectivity of the design was limited, would the express design perform better? Another hypothesis worth exploring is that there may be a "sweet spot" for the express methods in the proportion of items shown per block. The express design used in this study only showed 30% of the total items tested, but if that proportion were increased, possibly to 50%, different conclusions may be drawn.

Mike Serpetti

Claire Gilbert

Megan Peitz

## REFERENCES

Chrzan, K. and M. Patterson (2006) "Testing for the Optimal Number of Attributes in MaxDiff Questions," 2006 Sawtooth Software Conference Proceedings, 63–68.

Fairchild, K., B. Orme and E. Schwartz (2015) "Bandit Adaptive MaxDiff Designs for Huge Numbers of Items," 2015 Sawtooth Software Conference Proceedings, 105–117.

Finn, A. and J. J. Louviere (1992), "Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety," Journal of Public Policy and Marketing, 11, 12–25.

Hendrix, P. and S. Drucker (2008) "Alternative Approaches to MaxDiff with Large Sets of Disparate Items—Augmented and Tailored MaxDiff," 2007 Sawtooth Software Conference Proceedings, 169–187.

Horne, J. B., R. Rayner and S. Lenart (2012), "Continued Investigation into the Role of the 'Anchor' in MaxDiff and Related Tradeoff Exercises," 2012 Sawtooth Software Conference Proceedings, 43–58.

Lattery, K. (2011) "Anchoring Maximum Difference Scaling Against a Threshold—Dual Response and Direct Binary Responses," 2010 Sawtooth Software Conference Proceedings, 91–106.

Louviere, J. J., D. Street, L. Burgess, N. Wasi, T. Islam and A. A. J. Marley (2008) "Modeling the Choices of Individual Decision Makers by Combining Efficient Choice Experiment Designs with Extra Preference Information," Journal of Choice Modeling, 1: 128–63.

Orme (2005) "Accuracy of HB Estimation in MaxDiff Experiments," Sawtooth Software Research Paper available at https://sawtoothsoftware.com/support/technical-papers/maxdiff-best-worst-scaling/accuracy-of-hb-estimation-in-maxdiff-experiments-2005.

Orme, B. (2006) "Adaptive Maximum Difference Scaling," Sawtooth Software Research Paper available at https://sawtoothsoftware.com/support/technical-papers/maxdiff-best-worst-scaling/adaptive-maximum-difference-scaling-2006.

Wirth, R. and A. Wolfrath (2012) “Using MaxDiff for Evaluating Very Large Sets of Items,” 2012 Sawtooth Software Conference Proceedings, 59–78.

# NAÏVE BAYES CLASSIFIERS, OR HOW TO CLASSIFY VIA MAXDIFF WITHOUT DOING MAXDIFF

*DAVID W. LYON*
*AURORA MARKET MODELING*

## INTRODUCTION

MaxDiff utilities are a popular basis for segmentation, in part because they avoid the scale-usage bias often seen with batteries of ratings scales. But building a classifier (or "typing tool") for MaxDiff-based segments can be a problem. Why? Because typing tools usually do best when built using the original basis variables, but with MaxDiff, the utilities are not available to the classifier.

Classifiers in general (not just for MaxDiff) need to use as few questions as possible; we do not want to repeat a whole original questionnaire, and seldom even want to repeat every one of the original basis variables. Also, many typing tools need to work "on the fly," so respondents can be assigned to different concept tests, focus groups or quotas as soon as they are screened. In the case of MaxDiff, this means we don't want to use as many tasks as the original questionnaire, or use all the original items, and we certainly can't wait for a full new sample on which to do a lengthy HB run or latent class run to get utilities. The utilities are great for segmentation, but we can't use them at typing-tool time.

This problem is widely recognized, and a number of "duct tape" practitioner approaches to get around it have been discussed. All, however, involve a degree of either violating the basic MaxDiff model or throwing away some information that is available. This is unfortunate, because Bryan Orme and Rich Johnson (2009) published a *Marketing Research* magazine article that showed a theoretically sound solution that fully respects the MaxDiff model.

The method Orme and Johnson used is called Naïve Bayes Classification ("NBC"). Their article applied NBC to the MaxDiff situation, but did not discuss or name the underlying NBC method they were using. This paper has two main objectives: first, to further popularize the Orme and Johnson idea, and second, to introduce Naïve Bayes Classification as a general and widely useful approach to classification. NBCs can be a useful alternative to the usual discriminant analysis, multinomial logit regression or tree-based methods (CART, random forests, etc.).

Others have addressed the MaxDiff classification problem in other ways. Jay Magidson (2016) showed an elegant and easy solution, not using NBCs, when the typing tool is meant to use only *pairs* of items. It can be extended to more typical MaxDiff tasks of triples, quadruplets, etc., but much less easily as non-standard software will then be needed. Lech Komendant (2016) shows, in the next paper in this volume, other approaches not only for MaxDiff but also for ordinary choice tasks.

## NAÏVE BAYES CLASSIFIERS

This section introduces the general ideas behind Naïve Bayes. This discussion is independent of MaxDiff; how to embed MaxDiff into NBCs will be the topic of the next section.

Let’s begin with a trivial example. Suppose we have segments, let’s say there are three, and are asked to classify them using only a single variable, let’s say a four-level region. What can we do? Not much, but we could look at a crosstab like this

***Classifying from One Categorical Variable***

| | Counts | | | | Row Percentages | | |
|---|---|---|---|---|---|---|---|
| | **Seg A** | **Seg B** | **Seg C** | | **Seg A** | **Seg B** | **Seg C** |
| Northeast | 52 | 31 | 40 | | 42.3% | 25.2% | 32.5% |
| Midwest | 85 | 50 | 42 | | 48.0% | 28.2% | 23.7% |
| South | 87 | 89 | 43 | | 39.7% | 40.6% | 19.6% |
| West | 86 | 77 | 55 | | 39.4% | 35.3% | 25.2% |

and use it to classify. If we know a respondent is in the South, let’s say, we can find the largest entry in the South row, in this case 89, and conclude that our best bet is to classify them into Segment B. More precisely, we could look at row percentages for the South row, and say that there is a 40.6% chance the respondent is in Segment B, and a nearly equal chance he or she is in segment A.

Note that we are using row percentages when we do this, while the usual use of a crosstab like this in a segmentation context would focus on column percentages. The column percents are used in profiling the segments, to evaluate them and describe them to the client. We can relate the two percentages by observing that the segment sizes (at the total sample level) times a row of column percentages (for any region’s row), if re-percentaged by dividing all those products by the region size, produce the row percentages.

Thus we can say that the estimated probabilities of a respondent being in each segment are the vector of *column* percents for that respondent’s region, times the vector of overall segment sizes, re-percentaged (i.e., divided by their total so that the total becomes 100%; their original total happens to be the overall percentage of the sample in that region). See the table below for concrete details, continuing our region example.

The reason for separating the row percentages into column percentages and segment sizes is that we need the two split apart when we look at using multiple variables to classify, rather than using a single variable.

***Separating Row Percentages into Segment Sizes and Column Likelihoods***

| | ***Counts*** | | | **Reg** | | ***Row %s*** | |
|---|---|---|---|---|---|---|---|
| | **Seg A** | **Seg B** | **Seg C** | **Sizes** | **Seg A** | **Seg B** | **Seg C** |
| Northeast | 52 | 31 | 40 | 16.7% | 42.3% | 25.2% | 32.5% |
| Midwest | 85 | 50 | 42 | 24.0% | 48.0% | 28.2% | 23.7% |
| South | 87 | 89 | 43 | 29.7% | 39.7% | 40.6% | 19.6% |
| West | 86 | 77 | 55 | 29.6% | 39.4% | 35.3% | 25.2% |
| **Seg Sizes** | 42.1% | 33.5% | 24.4% | | | | |

| | ***Column %—Likelihoods*** | | | | | | |
|---|---|---|---|---|---|---|---|
| Northeast | 16.8% | 12.6% | 22.2% | Sizes | 42.1% | 33.5% | 24.4% |
| Midwest | 27.4% | 20.2% | 23.3% | West Col %s | 27.7% | 31.2% | 30.6% |
| South | 28.1% | 36.0% | 23.9% | Size X Col % | 11.7% | 10.4% | 7.5% |
| West | 27.7% | 31.2% | 30.6% | ÷ West reg size | 39.4% | 35.3% | 25.2% |

When running through these mechanics, we are simply applying Bayes' Rule to the classification process. The priors are the segment sizes (i.e., the probability of membership in each segment before considering *any* respondent data), the likelihoods are the row of column percentages (i.e., the likelihood of the region given the segment) and the posterior probabilities are the row percentages (i.e., the classification probabilities we seek).

***Bayes' Rule:***

Pr ( Seg X | Reg G ) = Pr ( Seg X ) × Pr ( Reg G | Seg X ) / Pr ( Reg G )

"Posterior" = "prior" × "likelihood" / "evidence"

Posterior ∝ prior × likelihood

Segment probabilities = Tabled Row Percents (from counts)

Segment probabilities ∝ Segment Sizes × Column Percents

It's obvious that our simple one-variable classifier will work poorly; not only is region a poor predictor in this case, but classifiers invariably need at least a handful of variables, if not a couple of dozen, even with good predictors. How can we add more variables here?

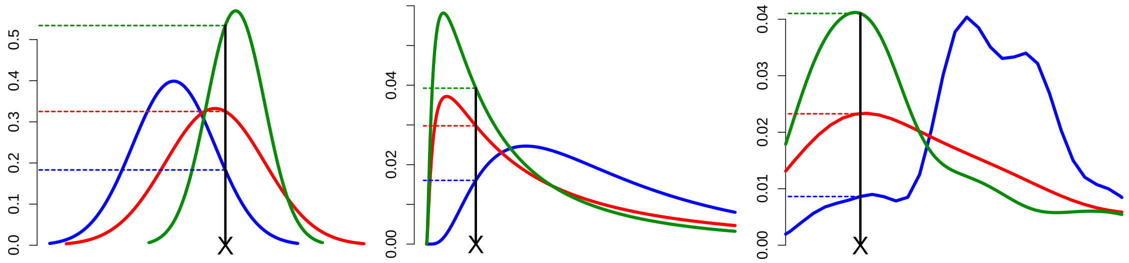
The process is simple. We calculate similar cross-tabs for all the other variables we want to use, focusing on the column percentages, and multiply the segment sizes (prior) times all the applicable rows from the various crosstabs (the likelihoods of the data given a segment) for all the variables. If a respondent is in the West, college-educated, high-income and female, we find the right row in each of four crosstabs for region, education, income and gender, multiply them all together along with the segment sizes, and re-percentage. In the example below, the respondent is 62.6% likely to be in segment B.

***Multiplying Prior and Likelihoods to Incorporate Multiple Variables***

| | **Seg A** | **Seg B** | **Seg C** |
|---|---|---|---|
| Priors (segment sizes) | 42.1% | 33.5% | 24.4% |
| *Variable: Value* | *Likelihoods (from column percents)* | | |
| Region: South | 27.7% | 31.2% | 30.6% |
| Education: College | 32.0% | 49.0% | 19.0% |
| Income: $150K and up | 28.2% | 37.5% | 34.3% |
| Gender: Female | 40.0% | 60.0% | 55.0% |
| (Column) Products | 0.00421 | 0.01152 | 0.00268 |
| Re-% to Probabilities | 22.9% | 62.6% | 14.5% |

What if a variable is continuous, not categorical? Instead of using crosstabs to get likelihoods (column percentages) for the data at hand, we can assume a distribution shape and use the probability density at the observed data point as the likelihood. We might assume normal distributions, using the observed mean and standard deviation for each segment, as illustrated in the leftmost plot below. If we don't like assuming normality, any other assumption can be used, as long as we can compute the probability density. If we want to avoid any assumptions at all, we can use an empirical density obtained from a kernel smoother in the same way.

**Likelihoods for Continuous Variables Are Their Probability Density Function Values**



*Assume normality* *Assume something else* *Empirical kernel smoothing*

In practice, it is common to bypass the problem and "bin" or categorize the continuous variables into, say, quartiles or quintiles or the like, and treat them like any categorical variable. It is also common to assume normality, sometimes after a transformation to make the data at least more symmetric or shorter-tailed. Many "continuous" survey questions are ratings on short scales (1 to 5, 1 to 7, 1 to 10, etc.) anyway; reducing those to categories is recognizing reality, not losing information.

NBCs rely on a very simple process and concept. We start with segment sizes as priors. We multiply them by a set of likelihoods for each variable in the classifier, with the likelihoods either a simple row of column percentages from a table or an easy evaluation of a probability density function for some distribution. We divide the final products by their total to get percentages summing to 100%, and assign a respondent to the segment with the highest final percentage. What could be simpler?

### Independence and Legitimacy

But, is this legitimate? If the variables are all independent of each other, then yes, it is a rigorous, defensible procedure. Of course, that is never true in practice, and often not even approximately true. Nevertheless, almost every introductory exposition of Naïve Bayes notes that it works “surprisingly well”! NBCs are a rare case of something that works in practice, but not in theory. The naïveté of assuming independence is what gives Naïve Bayes its name, and also alternative terms for it that include “Independence Bayes,” “Simple Bayes” and “Idiot Bayes”!

The independence assumption is not as strong, or as simple, as it first appears. The actual assumption required is not true independence at the total sample level, but *conditional* independence at the segment level. In everyday terms, we need to assume that the variables are independent *within each segment*. In principle, we can have conditional independence without overall independence.[1]

Does this matter in practice? It does, because one way of conceptualizing what latent class does is that it tries to find segments that maximize the conditional independence within segment. In other words, latent class is doing the best it can to give us conditional independence (on the original basis variables). How well it succeeds depends on how related the variables are, and on the number of variables and number of segments. But, it’s at least trying to create the situation that NBCs assume. Even if latent class was not used to find the original segments, *any* clustering approach must enhance conditional dependence to produce useful results, even if only as a by-product of whatever criterion it explicitly considers. So, yes, the fact that we only need conditional independence does make the independence assumption of NBCs more palatable, even if still far from true.[2]

In practice, we can avoid too much conditional dependence by our choice of variables to use. In an ordinary regression, adding a new variable that is highly collinear with the ones already in use helps predictive power very little, so we don’t. With NBCs, adding conditionally dependent new variables doesn’t help much, or even hurts, so again we don’t. More details on how we avoid that will follow.

## INCORPORATING MAXDIFF

How do NBCs help us classify using MaxDiff? We’ve seen how to incorporate continuous variables into NBCs, so an obvious first thought is to use MaxDiff utilities as variables in the NBC. But, as pointed out in the introduction, we can’t get those utilities in real time, and even if we could, we probably aren’t willing to ask as many tasks as would be needed to make them very accurate. Indeed, the secret of classifying via MaxDiff is to forget about using the utilities.

Instead, we observe that the answer to any single MaxDiff task is a categorical variable, and we know how to use categorical variables in an NBC. The answer to a MaxDiff task has two

[1] We can also have overall independence but not conditional independence, but examples of that are contrived and rare or non-existent in practice.

[2] Even conditional independence is not strictly required. At least in the case of normally-distributed variables, NBC has been shown to be an optimal classifier (in terms of hit rates) as long as each segment has the *same* conditional dependence patterns as the others (Zhang 2004). However, nothing in latent class or clustering methods promotes identical conditional dependence, so this is only a slight weakening of the assumption.

parts: the best item and the worst item. In this discussion, we will consider the two together. That means we will think of a 4-item task as producing one of 12 possible answers (one of 4 possible bests, and for each of those, one of 3 possible worsts among the remaining items). Each task's answer is a 12-level categorical variable.[3]

There is a problem. Any task we might be considering for a typing tool might not have been used in the original questionnaire, and almost certainly wasn't answered by all respondents. How then do we generate the crosstab from which to select one of the 12 rows of likelihoods (column percentages)? The basic idea is easy: if we have utilities for each segment, we can compute the likelihood of each answer from the multinomial logit model that underlies MaxDiff.

The formulas below give the details. They are standard multinomial logits, treating the best and worst answers as independent, except that they must eliminate (from the denominators) the impossible combination of the same item being both best and worst. Note that the easy computation in terms of scores works only for the score definition given below. The usual Sawtooth Software MaxDiff scores are adjusted in a way that does not work with the formulas shown.

***MaxDiff Response Probabilities from Utilities or (Unadjusted) Scores***

$$\Pr(\, a\; best,\; b\; worst \mid Task\,) \;=\; \frac{e^{U_a}\;\; e^{-U_b}}{\sum_{i,j\epsilon T, i\neq j} e^{U_i}\;\; e^{-U_j}} \;=\; \frac{e^{U_a - U_b}}{\sum_{i,j\epsilon T, i\neq j} e^{U_a - U_b}}$$

$$\propto\; Pr\,(\,a\; best\,)\;\; Pr\,(\,b\; worst\,)$$

$$\propto\; Score_a / Score_b \quad \text{where } Score_i = e^{U_i} \Big/ \textstyle\sum_{j=1}^{K} e^{U_j}$$

If our original segments came from latent class and we are keeping the "true" segments (meaning segments defined by a probability of inclusion for each respondent, as opposed to placing every respondent fully in their "modal" or most-likely segment), then the latent class model utilities are what we need. No matter where our segments came from, we can compute an *aggregate* MaxDiff model for each segment, and use those utilities for this purpose.

But what if we have individual respondent-level utilities estimated by hierarchical Bayes ("HB")? An obvious idea is to average the utilities for each segment and go from there. But, that approach leads to suboptimal performance. A better procedure is to use the MaxDiff model to transform the individual-level utilities into individual-level probabilities of each response to a task, and then average those *probabilities* (not the utilities) for each segment.

This is one of numerous situations in statistics where it is best to keep all transformations at the lowest level (i.e., the respondent level in this case) as long as possible before averaging up or otherwise aggregating. In this case, the key is that we want to average information that is already in the terms we want (probabilities of response, as in a crosstab) rather than average underlying ideas like utilities and then transform them. The author has verified that this procedure performs

[3] Another common way of thinking about a task's answer is that it produces two variables, not one, each having as many levels as there were items in the task. We could use that thinking instead, with no difference in any conclusions or outcomes, as long as we always use both the variables, or neither, from a given task.

more accurately than transforming averaged utilities and that it almost exactly duplicates the results from building an aggregate MaxDiff model for each segment (details are beyond the scope of this paper).

## IMPLEMENTATION PRACTICALITIES

At this point, we know what NBCs are and have seen how we can use answers to MaxDiff tasks with them. But a number of practical details must be dealt with:

- How do we know which tasks to use?
- If the answer is "those that work best" (and it will be), how do we evaluate that?
  - How can we tell from our original sample data?
  - What's the criterion for how well a task works?
- How can we decide which ordinary (non-MaxDiff) questions to use?
  - When should we use ordinary variables, vs. using MaxDiff tasks?
- How many tasks/variables is "enough" (or too many) in the NBC?
- How can we honestly estimate the expected accuracy of the final NBC?

### Which Tasks to Use

We know how to use MaxDiff tasks for classification, but *which* tasks should we use? The answer is simple: the ones that work best. Much as we do when using stepwise discriminant analysis to build a typing tool, we can conduct a stepwise search. First, we find the best task to use on its own, then the one that helps the most given that we already have the first one, then the third one to add to the first two, etc.

But, we may not want or be able to evaluate every possible task. With 25 items, there are 12,650 possible quadruplets to choose from, and far more if we want to consider quintuplets or if we have more items. We may not want to evaluate all the possibilities. Instead, we will use a random subset of tasks as starting points and apply a greedy algorithm to try to improve them, taking the best final product.

First, note that if there are few items, or if we want to use triplets or pairs in our typing tool tasks (and pairs may be an attractive option for typing tools to be administered by phone), the numbers are small enough that we can evaluate every possibility after all. When that's feasible, we do that and pick the best.

With larger problems, the author's strategy has been to generate 500 random tasks out of all the possibilities, evaluate them and choose the 10 that perform best. Then, each of those 10 is improved as much as possible by simple item swaps until no further improvement is possible. Whichever of the 10 improved finalists does best is then accepted as the one to use. This process is not guaranteed to find the best possible, but practical experience suggests that it often does (it is common to find many, or all, the 10 best starts converging to the same final task, for example). Nor is the best possible necessary; we need to be good, not perfect.

The greedy improvement strategy is to consider all possible single-item swaps. If we have 25 items in total and are considering quadruplets, then we have 4 items in the quad we are considering and 21 not in it. Any of the 4 could be replaced by any of the 21, so there are 84 possible single-item swaps. We evaluate each of the 84, and take whichever of them does best.[4] We repeat that process on the new quad until none of the 84 is better than the current quad. At that point, the algorithm is done; it has improved that quad as much as possible given the "greedy" simplification of looking at only one swap at a time.

Note that we are free to consider any size task in the typing tool. If the original questionnaire used quadruplets, we can still use triples, or pairs, or quintuplets, or of course quads, in the typing tool. Any of those "works" in the sense of the basic theory underlying MaxDiff and NBC. Many (this author included) have an automatic preference for keeping the same task size as in the original questionnaire, but this is by no means required and should not be automatic. We are free to consider the practicalities of typing tool implementation (which may argue for smaller tasks), and to experiment to see how well we can do with different task sizes.

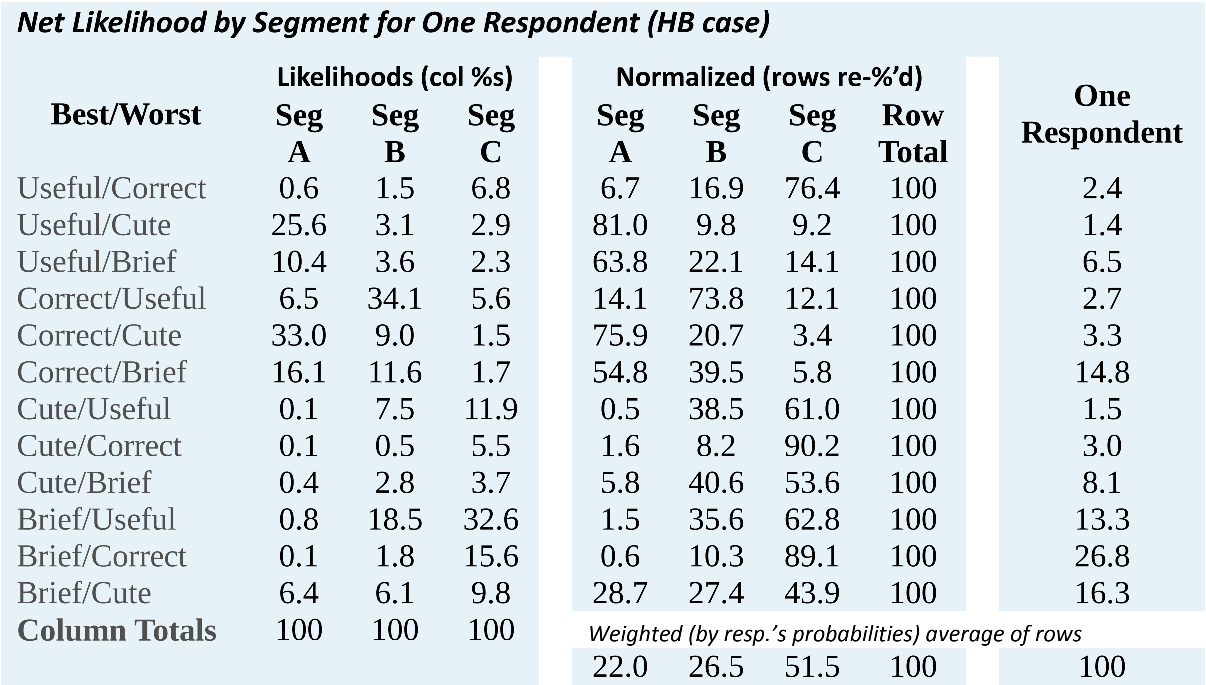
Also note that we are free to exclude a few MaxDiff items from consideration when searching for tasks to use in the classifier. Doing so does not disrupt the underlying theory in any way. Of course, it may hurt classification accuracy if we exclude an item that is viewed very differently across the segments!

### Evaluating What Works from the Original Sample

We want the tasks that work best, but how do we evaluate how well any given task works? We again have the problem that any task we might be considering probably wasn't answered by all respondents, perhaps not by any at all. The solution is, in spirit, the same as before: use the utilities we do have to determine how any given respondent would have answered the task.

First, consider the case where we have HB-estimated individual utilities. They give us a probability of each response for each respondent. This is, of course, different from the situation for ordinary questions where we know, with certainty, which one answer the respondent chose. We could assume the one MaxDiff task response with the highest probability, but that would ignore the uncertainty in the possible answers. Instead, we use the probabilities of each possible answer as weights, applied to the segment tables of likelihoods of each answer by segment, to obtain the likelihoods that would apply to that respondent, as illustrated by the table below.

[4] This process is a simple special case of the "modified Federov swaps" that are used in most algorithms for finding d-optimal experimental designs.



***Net Likelihood by Segment for One Respondent (HB case)***

| | Likelihoods (col %s) | | | Normalized (rows re-%'d) | | | | |
|---|---|---|---|---|---|---|---|---|
| **Best/Worst** | **Seg A** | **Seg B** | **Seg C** | **Seg A** | **Seg B** | **Seg C** | **Row Total** | **One Respondent** |
| Useful/Correct | 0.6 | 1.5 | 6.8 | 6.7 | 16.9 | 76.4 | 100 | 2.4 |
| Useful/Cute | 25.6 | 3.1 | 2.9 | 81.0 | 9.8 | 9.2 | 100 | 1.4 |
| Useful/Brief | 10.4 | 3.6 | 2.3 | 63.8 | 22.1 | 14.1 | 100 | 6.5 |
| Correct/Useful | 6.5 | 34.1 | 5.6 | 14.1 | 73.8 | 12.1 | 100 | 2.7 |
| Correct/Cute | 33.0 | 9.0 | 1.5 | 75.9 | 20.7 | 3.4 | 100 | 3.3 |
| Correct/Brief | 16.1 | 11.6 | 1.7 | 54.8 | 39.5 | 5.8 | 100 | 14.8 |
| Cute/Useful | 0.1 | 7.5 | 11.9 | 0.5 | 38.5 | 61.0 | 100 | 1.5 |
| Cute/Correct | 0.1 | 0.5 | 5.5 | 1.6 | 8.2 | 90.2 | 100 | 3.0 |
| Cute/Brief | 0.4 | 2.8 | 3.7 | 5.8 | 40.6 | 53.6 | 100 | 8.1 |
| Brief/Useful | 0.8 | 18.5 | 32.6 | 1.5 | 35.6 | 62.8 | 100 | 13.3 |
| Brief/Correct | 0.1 | 1.8 | 15.6 | 0.6 | 10.3 | 89.1 | 100 | 26.8 |
| Brief/Cute | 6.4 | 6.1 | 9.8 | 28.7 | 27.4 | 43.9 | 100 | 16.3 |
| **Column Totals** | 100 | 100 | 100 | *Weighted (by resp.'s probabilities) average of rows* | | | | |
| | | | | 22.0 | 26.5 | 51.5 | 100 | 100 |

Conceptually, we are using the expected likelihood for the MaxDiff response, with the expectation over all possible answers. Mechanically, instead of choosing a single row of likelihoods (like we would by picking a single region row out of the region crosstab), we take a weighted average of all the rows in the MaxDiff likelihoods table, weighted by the chances of each row being that respondent's response. This is one major point of distinction vs. many ad hoc approaches—we do not presume to "know" one correct answer for each respondent, but account for all the possibilities.

What about latent class-based segments where we don't have individual-level utilities? Here, we use the segment-level utilities to generate likelihoods of each response and assume they are the same for all respondents in the segment. If we are using modal segments (i.e., each respondent is assigned only to the segment for which his latent class likelihood was highest), those segment-level utilities are from an aggregate MaxDiff model for each segment, estimated after the fact. In effect, we are using one column of the task-by-segment likelihood tables—the one for the segment a respondent is in—to weight together the rows of the table to get a single vector of likelihoods for the respondent.

If we are using "true" segments, we don't know for certain what segment a respondent is in. In that case, we use the segment probabilities from latent class to weight together the likelihoods for each possible segment. In other words, we weight the rows of the task-by-segment likelihood table together once for each segment, using that segment's column of the table as weights, and then weight those results by segment together based on the segment likelihoods.[5]

[5] Magidson (2016) avoids this by simulating answers instead, an approach that is simpler, but less exact due to the random nature of simulation, and which this author has not explored.

## Criterion for What Works

How shall we decide which task or variable works best? An obvious way is to look at the hit rate when it is included. However, there are problems with hit rates, in any classifier, and special issues with NBCs. A general problem is that hit rates are “chunky”: they take a bump up when one extra respondent is classified correctly, but do not budge when any improvements in the underlying probabilities are not quite enough to push one more respondent “over the edge” into the correct segment. They are simply not sensitive or precise.

To see the particular problem with hit rates in NBCs, let’s look at what happens when we violate the key NBC assumption of conditional independence. Consider the most extreme possible dependence: what if we include the exact same variable several times? The table below shows the results as we repeatedly include an education variable, for a respondent with a post-grad education (ignoring segment sizes).

***Repeated Inclusion of Highly (Perfectly) Dependent Variables***

| | One-Variable Likelihoods | | | Net Probabilities Up to This Variable | | | Classification Result | In-Sample | |
|---|---|---|---|---|---|---|---|---|---|
| **Variable: Value** | **Seg A** | **Seg B** | **Seg C** | **Seg A** | **Seg B** | **Seg C** | | **Hit %** | **RLH** |
| Educ: Post-grad | 0.32 | 0.49 | 0.19 | 0.32 | 0.49 | 0.19 | B | 0.49 | 0.357 |
| Educ: Post-grad | 0.32 | 0.49 | 0.19 | 0.27 | 0.63 | 0.10 | B | 0.49 | 0.337 |
| Educ: Post-grad | 0.32 | 0.49 | 0.19 | 0.21 | 0.75 | 0.04 | B | 0.49 | 0.290 |
| Educ: Post-grad | 0.32 | 0.49 | 0.19 | 0.15 | 0.83 | 0.02 | B | 0.49 | 0.234 |

*Confidence in B is exaggerated; hit rate is not affected; RLH raises the alarm.*

The likelihoods for post-grad respondents, re-percentaged to add to 1.00, tell us 49% of them are in segment B, which is the most likely segment, so we get a hit rate of 49%. So far, so good. If we add the variable again, the process of multiplying likelihoods for different variables squares our original likelihoods. After re-percentaging, the segment B likelihood is now 0.63. That is still largest, so we still put all these respondents in segment B and are still right 49% of the time. If we add the same variable a third and fourth time, the segment B probability climbs to 0.83, staying the highest, so the hit rate stays at 49%. We have exaggerated our confidence about placing these respondents in segment B, and completely violated the conditional independence assumption, without a hint of trouble in the hit rate![6] Indeed, in data science literature, the surprising success of NBCs is often explained in terms of their hit rates being insensitive to various errors.

A more sensitive and useful criterion is the root likelihood or “RLH.” It is the geometric mean[7] of the probabilities assigned to the correct segments for the respondents. RLH will be familiar to many readers in the context of choice experiments, where it is commonly computed at the respondent level, as the geometric mean across the various choice tasks that respondent answered. Its use here is different; we have one classification probability per respondent and are taking the geometric mean across respondents, not within each one.

[6] In fairness, the hit rate did not go up and actually encourage this either.

[7] Where a regular average or “arithmetic mean” is the *sum* of all the values divided by *n*, the geometric mean is the *n*th root of the *product* of all the values, or the product raised to the *1/n* power. The computation process in practice is to take the exponential of the average of the natural logarithms of the values. If there are weights, they can be applied in the usual way when the logs are averaged.

In the table above, we see the RLH (computed here for the post-grad education respondents, but similar patterns would apply in the total sample) declining, faster and faster, as we make the mistake of introducing a highly (in this example, perfectly) dependent variable over and over. In effect, the RLH is penalizing our exaggerated confidence for segment B as we go. For the 19% of these respondents in segment C, we are assigning a probability of 0.02 to the correct segment, and for the 35% in segment A, a 0.15 correct probability. In the geometric mean, these low values far outweigh the (illegitimate) gains from claiming 0.83 certainty for the 49% in segment B.

Consequently, our key criterion in the stepwise search process will be the RLH. It is sensitive to the details of the classification probabilities a typing tool produces, not just to the final assignment and the hit rate.

### Including "Ordinary" Variables

It is important to understand that an NBC treats ordinary survey questions and MaxDiff tasks in the same way. Many early readers of this paper and the conference presentation interpreted survey questions as "covariates," or as providing a starting point from which we will improve by adding MaxDiff tasks. But there is no need to view them differently from the MaxDiff tasks. They are not *co*-variates, they are simply other variables.

This means that we can search through the available survey variables in a stepwise fashion, much as we do for MaxDiff tasks, and indeed that is what the author would recommend in most cases, and what he did in the second case study presented below.

Of course, we may want to exclude certain questions from a typing tool (many clients are reluctant to ask for income in a short consumer-oriented questionnaire, for example). As in any other classification situation, the original basis variables tend to predict much better than non-basis variables. Thus, we may decide to simplify or speed up our search by considering only basis variables.[8] If MaxDiff was the only basis for segmentation, we might not consider other survey questions at all. Or, we might consider only those survey questions that will be needed in a new screener anyway, whether basis variables or not, making them "free" to the typing tool.

There is no inherent theoretical or statistical reason to avoid any survey questions, or to force any into an NBC. They can be excluded for practical reasons, but otherwise considered for inclusion in competition against possible MaxDiff tasks, with the ones that help the most being included.

### Significance Testing in the Stepwise Search

As we build an NBC stepwise, we need a criterion for when to stop. This may be as simple as stopping when the in-sample hit rate or RLH is "high enough," but if we have high standards, that may never happen. At some point, additional tasks or variables aren't adding enough to be legitimate and we cross the line into overfitting. Significance testing helps prevent this.

In addition, when considering survey questions with varying numbers of categories, and MaxDiff tasks with larger numbers of possible answers, we don't want to always favor questions

[8] If some basis variables are excluded from the typing tool, it may be much more worthwhile to search for other survey questions that might help substitute for them; a wider search would be appropriate in that case.

with more levels just because it is easier for them to increase RLH. Significance testing is a way to put all possible items on a more equal footing; we can choose the most significant option rather than the one with highest raw RLH increase.

We can test significance using a Likelihood Ratio Test ("LRT"). This test involves comparing the log-likelihood for a model vs. that for the same model with one more variable or task added. Twice the difference in log-likelihoods (the difference in logs is the log of the *ratio* of the two likelihoods; hence the name) is distributed as a chi-squared under the null hypothesis of no contribution from the new variable. The log-likelihoods are easy to compute: they are *N* times the natural log of the RLH, where *N* is the number of respondents.

For survey questions, that chi-squared is on *m-1* degrees of freedom if the variable has *m* possible levels. The appropriate degrees of freedom is not clear-cut in the case of MaxDiff tasks. We could say that a quadruplet leads to 12 different possible answers and use 11 d.f. in that case. But then we are ignoring the fact that those 12 are structured from only 4 underlying utilities. In view of that, the author recommends the number of items in the MaxDiff task as the d.f. for the chi-squared.[9] He is not certain that this is the correct procedure, however, and would appreciate any feedback. When we choose what to add next to an NBC, the assumption made here is critical in choosing between the best possible MaxDiff task and the best available survey question; it is an important question.

A critical part of the significance testing is deciding at what level we will stop. When we are testing tens or hundreds of thousands of possible MaxDiff tasks, and perhaps hundreds of survey questions, we cannot simply apply a 0.05 significance criterion to each one. Doing that would lead to dozens or hundreds of variables being added even when all were generated randomly.

There are many ways to deal with this multiple testing issue. The author uses the simple, but highly conservative and stringent, Bonferroni procedure. Its basic idea is that if we want an overall "experiment-wise" significance level of 0.05, we should require any one variable to meet a criterion of *0.05/k*, where *k* is the number of variables we are considering. If we have 250 survey variables, for instance, we would use only the ones that pass a 0.0002 level of significance.

As with the d.f. in the chi-squared, it is not clear-cut how to apply this idea to the MaxDiff tasks. Do we say that there are 12,650 possible quadruplets of 25 items we are considering and ask for 0.05 / 12,650 = 0.000004 as our criterion? That is extreme, since there are only 25 underlying utilities generating all those options. The author's practice has been to use the total number of MaxDiff items as the number of variables being tested. This may be too lenient, and is another topic where feedback would be welcome.

An easy way to implement the Bonferroni process is to calculate a conventional significance level for each variable or task, and then multiply it by the number of variables being considered, and compare that to the desired overall level (such as 0.05). (In other words, we can adjust the significance level for each variable by multiplying, rather than adjusting the criterion to be met by dividing.) If we do that both for the survey variables and for the MaxDiff tasks, we then have them on a common significance scale and can decide which one helps the classifier the most. If

[9] In the conference presentation, and in the implementation of the Case Studies in a later section, the author used the number of items in a task minus one. Further thinking since then has convinced him that "minus one" is not appropriate. All the item utilities for a task are "free" parameters. Since none are constrained by the others, all should be counted toward degrees of freedom.

the MaxDiff correction is in fact too lenient, this process will favor MaxDiff tasks unduly. Judging from the face validity of the results in Case Study 2 (in a later section), this does not appear to be occurring. The Bonferroni correction for MaxDiff tasks in that case seems not to be lenient by a large margin, if at all.

We could take all this one step further and "penalize" questions for their complexity by further adjusting the significance calculations. A MaxDiff task, for example, involves two questions and might be viewed as a bigger interviewing burden than an ordinary survey question. Or, a ratings item from the survey might be viewed as more costly if no other items using its scale have yet been added to the classifier, but less costly if the necessary scale and question introduction are already necessary for earlier items. The goal would be some kind of "RLH bang for the buck" criterion that reflects both classification improvement and interviewing burden. One obvious idea would be to divide the RLH gain by some cost measure (that would be set at 1.0 for most questions), before applying the LRTs.

### Evaluating the Final Product

Any stepwise procedure involves a degree of capitalizing on chance. We control this somewhat by careful significance testing. But to honestly estimate the performance of the final classifier, we need some type of out-of-sample testing.

The classic market research response to this is the holdout sample. Modern data science practitioners go one step further, dividing their sample in three parts: a "training" sample from which to estimate parameters (i.e., in this case to generate the simple crosstabs and the MaxDiff utilities at the segment level), a "validation" sample that drives decisions as to which tasks and variables to use (in our case, RLHs would be computed on the validation sample as we do the stepwise search) and a "test" sample, like our usual holdout samples, on which to gauge the final performance.

But market research samples are often smallish, and when divided into three parts, the validation and test samples may be too small to provide stable estimates of performance. This author would much rather use an approach that keeps the entire sample together.

The author uses an approach known to statisticians as "jackknifing" and to data scientists, more descriptively, as "leave one out" or "LOO." In it, we pick a single respondent, and leave him or her out while computing all the model likelihoods and tables on the rest. Then we calculate segment classification probabilities for the one left out, based on all the rest, and see how we did. That process is repeated for each of the original respondents. Each is predicted only by the others, creating "N holdout samples of one." This is not the computational burden it seems; it can be fast with careful programming.

This procedure is honest in terms of the likelihood estimates, and does not give credit for overfitting in estimating them. What it does not account for, however, is that the original stepwise decisions of which tasks and variables to use might have been different had a given respondent been left out of the original search. In principle, we could conduct the entire stepwise process for each respondent, but that would indeed be a crushing computational problem. Applying LOO to the likelihood estimates is a major step toward an honest estimate of overall performance, and it is practical to implement. We use it even though it does not quite cover all the issues.

It is crucial that the LOO assessment be applied *once*, to the one final classifier, after all other decisions have been made. If we were to apply LOO along the way to help decide what to do, we would be converting our left-out respondents from a test sample to a validation sample, and be back to overfitting. In the case studies presented here, LOO evaluations are shown for several alternative models. To be clear, the point of that is not to suggest looking at them and choosing the best. They simply indicate what the final assessment would have been for each of several different possibilities that might have been decided on based on in-sample statistics.

### Summarizing the Implementation

In sum, our process for building a Naïve Bayes Classifier stepwise will be:

- Generate 500 possible MaxDiff tasks, pick the 10 best, apply greedy improvements to each, and choose the best final one of the 10 results. Calculate its significance on a Bonferroni-corrected LRT basis.
- Evaluate each survey variable (if they are being considered) via a Bonferroni-corrected LRT and choose the one with the best final significance.
- Stop the stepwise search if neither the best task nor the best survey question passes our overall significance criterion.
- Add the best task or the best variable, whichever had the lower corrected significance level, to the classifier so far, and repeat from the top.
- Examine in-sample RLH and hit rates and decide whether to stop earlier than the significance tests suggested (based on practical issues).
- Evaluate hit rates and RLH for the final classifier on an LOO basis.

## Case Studies

### Case Study 1: Textbook-Simple

The first case study is a straightforward MaxDiff-based segmentation. The MaxDiff exercise involved 14 items, with each respondent answering 11 quadruplets of items. That means each item appeared three times per respondent, in line with the most frequent recommendation for numbers of tasks. The items were characteristics of allergy medications. The 577 respondents were physicians treating allergies and the sample was unweighted. Individual-level utilities were estimated via HB, and then latent class was applied on the posterior means.[10] A 4-segment solution was chosen for the final segments. The MaxDiff utilities were the only basis variables; no other questions were used.

To build the NBC, we started with quadruplets, as in the original exercise, and excluded all other survey questions (since none were basis variables anyway). As detailed in the table below, significance testing stopped adding tasks after 9 were used, and produced an NBC with a hit rate over 90% (both in-sample and LOO). From a practical point of view, the last two tasks added

[10] While this is a common procedure in commercial practice, it can't be recommended on theoretical grounds. See Eagle (2013) for one discussion of why not. As Jay Magidson has said, this process amounts to "assuming there are no segments [by adopting HB], then trying [via LC] to find those segments that don't exist." Nevertheless, it is in wide use.

pushed the RLH up only a little and the hit rate hardly at all. We can still hit 90% with 7 tasks. We had 11 tasks in the original exercise; neither 9 nor 7 feel like much improvement. However, we can hit almost 85% with only 3 tasks, and 70% with just one! Those are outstanding results, with a mere 3 tasks achieving a performance better than most classifiers seen in practice.

***Case Study 1: Classification Accuracy Using Quadruplets***

| | # MD | In-Sample | | L-O-O | |
|---|---|---|---|---|---|
| **Classifier** | **tasks** | **RLH** | **Hits** | **RLH** | **Hits** |
| NBC, all significant tasks | 9 | 0.783 | 0.905 | 0.784 | 0.905 |
| NBC, cut off sooner | 7 | 0.751 | 0.899 | 0.751 | 0.901 |
| NBC, cut much sooner | 3 | 0.617 | 0.846 | 0.617 | 0.844 |
| NBC, cut ridiculously soon | 1 | 0.438 | 0.700 | 0.437 | 0.697 |

If we do that well with quadruplets, a natural question is whether we might get away with triplets or even pairs, resulting in an even less burdensome classifier. As shown in the table below, we can. Three triplets, or five pairs, either of which is a low respondent burden, will get us to 80% hit rates, a level that again beats most real-world classifiers. By the way, the largest segment in this study was 31% of the total—the high hit rates with minimal data were not achieved by putting almost everyone in a single huge segment.

***Case Study 1: Classification Accuracy Using Triplets or Pairs***

| | # MD | In-Sample | | L-O-O | |
|---|---|---|---|---|---|
| **Classifier** | **tasks** | **RLH** | **Hits** | **RLH** | **Hits** |
| NBC, all significant ***triplets*** | 10 | 0.766 | 0.905 | 0.767 | 0.901 |
| NBC, fewer triplets | 6 | 0.700 | 0.863 | 0.699 | 0.860 |
| NBC, still fewer triplets | 3 | 0.581 | 0.825 | 0.581 | 0.825 |
| | | | | | |
| NBC, all significant ***pairs*** | 13 | 0.710 | 0.877 | 0.709 | 0.877 |
| NBC, fewer pairs | 10 | 0.685 | 0.858 | 0.684 | 0.854 |
| NBC, still fewer pairs | 5 | 0.583 | 0.809 | 0.582 | 0.808 |
| NBC, ridiculously few pairs | 2 | 0.455 | 0.714 | 0.453 | 0.711 |

In this case study, the in-sample and LOO evaluations of performance, in terms of both RLH and hits, were close to each other. The LOO values were lower, as would be expected, but by slim margins. This suggests that the significance testing is preventing overfitting, and that the sample is large enough vs. the number of tasks being considered as to not facilitate overfitting.

## Case Study 2: Complex

The second case study involved *two* MaxDiff exercises, one with 27 items, the other with 26. Each concerned a software product; the two products were related but different. Twenty of the MaxDiff items were the same or similar between the two exercises, but of course applied to different products. There were 1047 small-business respondents, who each answered 11 quintuplets for the first MaxDiff, and 10 for the second, so each item was seen twice by each respondent. MaxDiff utilities were estimated via HB, and the MaxDiff posterior means were then combined with numerous survey questions via canonical correlation. Howard-Harris clustering on the canonical variates produced the final 6-segment solution. There were many moving parts here!

The stepwise search process to build the NBC considered both of the two MaxDiff exercises simultaneously, as well as almost all the available survey question variables. Many of the survey questions included were ratings on 1 to 5 scales, which were treated as continuous normally-distributed variables, not binned (an arguably sloppy, but still obviously effective, approach); about half those in the final classifiers were this kind of variable. The outcome of the search process is summarized in the table below. Note that the full final classifier, as determined by the significance-testing cutoff, involves more questions (16) and more MaxDiff tasks (nearly half the original total) than we would want to use in practice, although it does achieve about a 70% hit rate.[11] We can cut the number of variables and tasks in half and still get a hit rate (LOO) of 66%. The first 5 MaxDiff tasks and 3 survey variables bring us up to 60%, and 4 more survey variables add about 6% more.

***Case Study 2: Classification Accuracy Using Both MaxDiffs and Survey Variables***

| | # Survey Vars | #MD 1 tasks | #MD 2 tasks | In-Sample | | L-O-O | |
|---|---|---|---|---|---|---|---|
| **Classifier** | | | | **RLH** | **Hits** | **RLH** | **Hits** |
| NBC, all significant items | 16 | 4 | 6 | 0.491 | 0.725 | 0.438 | 0.703 |
| NBC, cut off sooner | 7 | 2 | 3 | 0.414 | 0.670 | 0.397 | 0.660 |
| NBC, cut much sooner | 3 | 2 | 3 | 0.376 | 0.605 | 0.366 | 0.600 |

Out of curiosity, we looked at NBCs using *only* MaxDiff tasks and *only* survey questions. Running them out until the final significance stop, we use 13 MaxDiff tasks but achieve only 55% hit rates, or 21 survey questions to reach only 57% on LOO hits (61% in-sample). Eliminating either big chunk of the original basis variables hurts; we need them both to classify well.

***Case Study 2: Classification Accuracy, MaxDiffs Only, Survey Questions Only, vs. Both***

| | # Survey Vars | #MD 1 tasks | #MD 2 tasks | In-Sample | | L-O-O | |
|---|---|---|---|---|---|---|---|
| **Classifier** | | | | **RLH** | **Hits** | **RLH** | **Hits** |
| NBC, all significant items | 16 | 4 | 6 | 0.491 | 0.725 | 0.438 | 0.703 |
| NBC, MaxDiff tasks only | 0 | 7 | 6 | 0.315 | 0.554 | 0.314 | 0.551 |
| NBC, no MaxDiff info | 21 | 0 | 0 | 0.346 | 0.607 | 0.313 | 0.574 |

The LOO evaluations are noticeably lower than the in-sample ones in this case study, reflecting the huge number of available tasks and variables and suggesting that the details of the LRT significance testing might in fact be too lenient.

## Summary

When segments are based on MaxDiff utilities, even if only in part, classifiers are unlikely to work well unless they can use MaxDiff results. But, that is not possible with most classification

[11] After the high rates seen in the first case study, 70% may seem low, but for a 6-segment solution of such complexity, and compared to many live commercial studies, it is quite reasonable and acceptable.

approaches. With Naïve Bayes Classifiers, however, incorporating MaxDiff tasks in the classifier is natural. We are free to use tasks of any desired size in the classifier, regardless of what task sizes were in the original study. We do need a search strategy to determine what tasks to use, and have presented one such strategy that works well in practice.

This procedure has key advantages over most of the alternatives that have been proposed. It fully accounts for what we know about each original respondent's likely answers to typing tasks; we use the probabilities instead of pretending certainty that the highest-utility item will be best, for example. Similarly, there are no assumptions along the lines of "Segment C will *always* choose item A as best." It treats both parts of the MaxDiff answers as a unified whole, using both the best and the worst responses. It uses actual MaxDiff questions (i.e., best and worst) rather than requiring a full ranking of all the items in a task. It does not attempt to create pseudo-utilities to substitute for the real ones. In short, it adheres fully to the fundamental MaxDiff multinomial logit model.

Naïve Bayes Classifiers, with or without MaxDiff, are another tool in the classifier toolbox; they can be considered alongside discriminant analysis, multinomial logit, CART, random forests, K-nearest neighbors, neural nets and a host of lesser-known (in market research circles) others. They are theoretically sound except for the (conditional) independence assumption. But they work much better in practice than might be expected, in part because judicious stepwise selection avoids inclusion of highly conditionally dependent variables. In machine learning circles, there are widespread claims that NBCs outperform other, more complex, options.

The basic ideas presented here can be extended to general choice models beyond MaxDiff (as Komendant 2016 illustrates), although segmentations based on general-purpose discrete choice are not as common as ones based on MaxDiff.

R includes several packages with functions to implement NBCs (e.g, *e1071* and *klaR*), but none that understand MaxDiff or implement the necessary stepwise search for tasks to use. However, after the original Orme and Johnson (2009) paper was published, Sawtooth Software implemented the basic process in a pair of programs designed for in-house consulting use, without a nice user interface. Bryan Orme has now offered to make those available, as is, to interested parties.

The Sawtooth Software programs do not use the exact search procedure described here for MaxDiff tasks, but achieve the same goal through similar means. The search program accepts ordinary survey question variables in addition to the MaxDiff, but the questions to be used must be specified in advance and are taken as given; they are not part of the stepwise search. One strategy to deal with this is to run a stepwise multiple linear discriminant analysis on the regular survey questions, and use its results to select the ones to include in the NBC via the Sawtooth Software programs.

David W. Lyon

## Appendix —"Smoothing" an NBC

Naïve Bayes Classifiers can run into problems when a category of one of their variables never occurs in a segment. If no one in segment 2 was ever in the West region, for example, every respondent in the West will have a zero likelihood for segment 2. No matter how many other variables are included, or how strongly they might all point to segment 2 for a respondent in the West, that zero gives region an absolute "veto" over that segment for the West. If there are many variables, and a number of zeroes amongst them, we might even have to classify a new respondent for whom *every* segment is forced to zero likelihood by one variable or another!

The fix for this is to ensure a minimum value in every cell. We can add 1.0 to every cell of a crosstab of counts before calculating percentages, a procedure known as "Laplace smoothing" (not to be confused with "Laplacian smoothing," which is unrelated). Or, we can add an arbitrary amount $\alpha$, with $\alpha$ typically less than 1 (perhaps 0.5), a generalization known as "Lidstone smoothing," where $\alpha = 1$ produces Laplace smoothing and $\alpha = 0$ is no smoothing. Or, we might simply raise any zero percentage up to some minimum.

Anything we do here has a distinct practitioner's "duct tape" feel to it. However, the Laplace and Lidstone "additive smoothing" approaches do at least have a degree of mathematical elegance: applying them yields percentages that are the Bayesian posterior mean estimates of the percentages under a symmetric Dirichlet prior with parameter $\alpha$. In effect, they are "shrinkage estimators" pulling all percentages toward a weak prior of equal probabilities for all categories of a variable.

Similar problems can occur with continuous variables if an outlier value on one variable produces a tiny probability density, and with MaxDiff tasks if an improbable answer is chosen. The author has controlled for that by imposing a minimum likelihood of 0.0001 (i.e., 0.01%) for all MaxDiff likelihood computations, and the same for all continuous PDFs.

## References

Eagle, Thomas (2013), "Segmenting Choice and Non-Choice Data Simultaneously," *Proceedings of the 2013 Sawtooth Software Conference*

Komendant, Lech (2016), "Typing Tools in the Context of Choice Experiments," *Proceedings of the 2016 Sawtooth Software Conference* (next paper in this volume)

Magidson, Jay and Gary Bennett (2016), "How to Develop a MaxDiff Typing Tool to Assign New Cases into Meaningful Segments," *Advanced Research Techniques Forum* presentation.

White paper at http://www.statisticalinnovations.com/wp-content/uploads/White-Paper-MaxDiff-Typing-Tool-FINAL.pdf

Orme, Bryan and Rich Johnson (2009), “Typing Tools That Work,” *Marketing Research*, Summer 2009. Available at http://www.sawtoothsoftware.com/download/techpap/typing_tools_mrmag.pdf

Sawtooth Software (2009), “Software for Typing MaxDiff Respondents,” *internal working paper.*

Zhang, Harry (2004), “The Optimality of Naïve Bayes,” Proceedings of the Seventeenth International Florida Artificial Intelligence Conference (“FLAIR”)

# TYPING TOOLS IN THE CONTEXT OF CHOICE EXPERIMENTS

***LECH KOMENDANT***
*IQS*

## INTRODUCTION

Choice data is a generally acclaimed excellent source for obtaining preference information. Especially MaxDiff enjoys the status of being almost the perfect method for preference elicitation in segmentation studies. CBC data could provide great segmentation material as well, even though usually conjoint studies are run for different purposes.

In many segmentation projects one of the main deliverables is a typing tool—a short length questionnaire that could be used for predicting segment membership outside the main segmentation study. Strategies for building such tools in choice based segmentation must embrace the fact that every person potentially had been given a different questionnaire. That is the reason why some people find it intimidating. But it does not have to be so.

In this article I will demonstrate that simple principles can help us to build an efficient typing questionnaire. And well known classification models can be used for segment prediction even in real time classification.

Building a typing tool for a choice-based segmentation study comes down to 4 interrelated parts:

- Decision on the type and number of questions,
- Selection of particular questions (feature selection),
- Selection of the classifying model, and
- Quality assessment.

I will elaborate on each of those elements during an analysis of the three basic approaches to typing tool construction. Those approaches were meant for MaxDiff-based segmentations but, as will be seen, they can also be successfully used in CBC and ACBC segmentations.

### Basic Typing Tools in Choice-Based Segmentations

I will describe two solutions for a typing tool problem using the MaxDiff data structure as examples. Then I will introduce a possible extension to canonical Naïve Bayes approach. Next, I will cover conjoint data structure as well. Three basic options are as follows:

> **Option 1:** Pairwise classifier as suggested by Thomas Eagle in 2012 on the Sawtooth Software LinkedIn group—this option lends itself to numerous implementations.
>
> **Option 2:** Extensive regression search suggested by Kevin Lattery in 2012 on the Sawtooth Software LinkedIn group.
>
> **Option 3:** Naïve Bayes + greedy search classifier tailored for MaxDiff (Orme & Johnson—Marketing Research 2009).

Options 2 and 3 represent specific combinations of the predictive model and the feature selection algorithm as used by different researchers. Of course, parts of those models can be replaced to a certain degree as we will see in the next part, namely a review of their classification efficiency.

## Pairwise Typing Tool

Let’s assume that we have run a simple MaxDiff study on the characteristics driving preferences for muffins. We used 6 items: fluffiness, moisture, flavor, crust, taste, color. We ran a clustering and got 2 segments: gourmands and aesthetes.

After HB estimation, a sample of our data will look like Table1:

**Table 1**

| ID | fluffiness | moisture | flavor | crust | taste | color | **Segment** |
|---|---|---|---|---|---|---|---|
| 1 | 3.46 | -2.09 | 4.28 | 2.42 | -0.32 | -0.41 | **1** |
| 2 | 2.05 | -3.94 | -0.97 | 0.89 | -0.19 | 1.12 | **2** |
| 3 | 2.85 | -3.53 | -1.85 | 1.98 | -0.34 | 0.13 | **1** |
| 4 | 1.19 | -4.26 | -1.34 | 1.50 | 0.48 | -1.12 | **1** |
| 5 | 0.85 | -4.15 | -1.52 | 3.06 | 0.99 | 2.01 | **2** |

The question format for a “Pairwise typing tool” is a Pairwise trade-off using the wording of the original question. Firstly, we have to identify which pairs work best for our tasks. To do this we put together all possible pairs of utilities (in our example we would only have 15 such pairs). For every person in each pair we decide using the First Choice rule the winning item:

**Table 2**

| pair 1 | | pair 2 | | pair 3 | | etc. |
|---|---|---|---|---|---|---|
| fluffiness | moisture | fluffiness | flavor | fluffiness | crust | *etc.* |
| 3.46 | -2.09 | 3.46 | 4.28 | 3.46 | 2.42 | *etc.* |
| 2.05 | -3.94 | 2.05 | -0.97 | 2.05 | 0.89 | *etc.* |
| 2.85 | -3.53 | 2.85 | -1.85 | 2.85 | 1.98 | *etc.* |
| 1.19 | -4.26 | 1.19 | -1.34 | 1.19 | 1.50 | *etc.* |
| 0.85 | -4.15 | 0.85 | -1.52 | 0.85 | 3.06 | *etc.* |

Next, we take each pair and make an indicator variable from it. After adding a segment membership flag (Table 3), it becomes our dataset on which we could run any possible classifier with any possible feature selection algorithm. I will not go into classifier and feature selection choice here. Later on, I will discuss some possible approaches. After selecting features, we should have a list of pairs which serves in our optimal classifier. It is based on those pairs that we will build our typing questionnaire.

**Table 3**

| Pair 1 | pair 2 | pair 3 | pair 4 etc. | **Segment** |
|---|---|---|---|---|
| 1 | 2 | 1 | etc. | **1** |
| 1 | 1 | 1 | etc. | **2** |
| 1 | 1 | 1 | etc. | **1** |
| 1 | 1 | 2 | etc. | **1** |
| 1 | 1 | 2 | etc. | **2** |

## Regression Using Rankings

Sometimes we might decide that using pairs is not efficient enough—based on the general MaxDiff experiments we know that using larger sets of trade-offs offers us more information and greater efficiency. We can use a very similar approach to the one given above to build a tool based on triplets, quads, quints, and so on. The only difference with the larger sets is that we do not build indicators of choice, but use sets of indicators of ranking instead (Tables 4 and 5). In our questionnaire, we can either use sets of rankings or best-worst exercises.

**Table 4**

| triplet 1 | | | *triplet 2* | | | *etc.* |
|---|---|---|---|---|---|---|
| Fluffiness | moisture | flavor | fluffiness | moisture | crust | *etc.* |
| 3.46 | -2.09 | 4.28 | 3.46 | -2.09 | 2.42 | *etc.* |
| 2.05 | -3.94 | -0.97 | 2.05 | -3.94 | 0.89 | *etc.* |
| 2.85 | -3.53 | -1.85 | 2.85 | -3.53 | 1.98 | *etc.* |
| 1.19 | -4.26 | -1.34 | 1.19 | -4.26 | 1.50 | *etc.* |
| 0.85 | -4.15 | -1.52 | 0.85 | -4.15 | 3.06 | *etc.* |

⬇

**Table 5**

| triplet 1 | | | triplet 2 | | | *etc.* | **Segment** |
|---|---|---|---|---|---|---|---|
| fluffiness | moisture | flavor | fluffiness | moisture | crust | *etc.* | |
| 2 | 1 | 3 | 3 | 1 | 2 | *etc.* | **1** |
| 3 | 1 | 2 | 3 | 1 | 2 | *etc.* | **2** |
| 3 | 1 | 2 | 3 | 1 | 2 | *etc.* | **1** |
| 3 | 1 | 2 | 2 | 1 | 3 | *etc.* | **1** |
| 3 | 1 | 2 | 2 | 1 | 3 | *etc.* | **2** |

The original approach, that Kevin Lattery suggested, used the stepwise algorithm with an exhaustive search on each step and logistic regression as classifier. That means that we have to:

1. Build all possible 1 triplet models,
2. Choose the best one on some criterion (e.g., minimal deviance),
3. Build all 2-triplets models with the first triplet set from the previous step,
4. Choose the best 2-triplets model. And so on.

For our simple 6-attributes study only 20 regressions are required for each step. But as the number of items grows, the number of models grows very fast as well. Given 20 items, 1,140 triplet regressions are required for each step. Except for the small problems, making all the required computations consumes a lot of time.

## Adaptive Naïve Bayes for MaxDiff

The typing tools we use are usually static—i.e., all people are given the same questions/ exercises and are classified according to the same model. On the other hand, in a "conjoint community," we are used to the notion of adaptiveness of questionnaires as a way to provide better results in a shorter time (and possibly a better experience for our respondents). The same premises stand behind an idea of the adaptive Naïve Bayes Classifier.

The concept is very simple: let us give a respondent a very short set of exercises to assess his vague segment membership. Having done that, we can proceed with giving him only exercises which probe his most probable segments. As this adaptive step is beyond what Sawtooth Software NB classifier does, we have to address certain issues at each step.

1. We have to use an adaptive questionnaire with real time computation of membership probabilities. It means that our survey software has to have this capability.
2. The preliminary part—which is the same for each person—should be efficient and unbiased. What this means is that we should optimize the model to have high overall accuracy and minimal differences between accuracies for segments.
3. The feature selection problem is "branched"—optimally we want to know which sets to show basing on both the probabilities of segment membership and the questions asked. It means that we have to prepare optimal questionnaires of a given length for many situations. Alternatively, we could use a simpler filtering strategy—e.g., choosing several

sets which differentiate the probability of the best-worst choices in the given pair of segments to the greatest extent.

4. The previous point becomes more complicated if we decide that we want to probe more than two segments—which could be a good idea if we have obtained many segments and possibly flat probabilities after the preliminary part.

From the perspective of the Naïve Bayes MaxDiff typing tool, the classification itself is straightforward. All we need to do is combine the preliminary and adaptive parts and compute likelihoods as with the use of the standard tool.

## EXTENSION FOR CBC STUDIES

Any of the above described tools can be used for typing CBC-based segmentation as well. One difference is, needless to say, the format of the questions—here they constitute discrete choices between profiles. The other is much more important though, namely it is the way of arriving at our optimal questionnaire.

In the conjoint HB data matrix we have utilities for every level of every attribute, but we will be showing full profiles to our respondents. The obvious solution would be to build all possible profiles from all of the attributes we have at our disposal, and then to proceed as if they were MaxDiff items. Unfortunately, it will only work for the simplest conjoint experiments. A Brand-Price-Pack study with 4 brands, 3 pack sizes and 5 price levels translates into 60 possible profiles. This is doable with the Pairwise classifier and Naïve Bayes (1,770 pairs). It probably will not be possible with triplet Regression on Ranking approach since there are 34,220 possible triplets of the full profiles. And those numbers rise very quickly making an all possible profiles strategy unfeasible in general.

For the Naïve Bayes tool we can leverage the strategy of greedy search. We simply add an additional layer to the search. We improve our randomly composed questionnaire swapping:

**Attribute levels**
  **In all attributes**
    In all profiles
      In all sets

CBC specific part of the loop

The criterion for improvement is the same, the classification model is also the same as for MaxDiff.

The same selection procedure can be used with the pairwise typing tool and the Regression on Rankings typing tool. Otherwise, you can pre-filter attributes/levels and choose those most discriminating of your segments and then use them to build a subset of possible profiles.

## TYPING TOOLS ACCURACY REVIEW

### General Notes

This section is devoted to the comparison of three specific approaches to building a typing tool for choice-based segmentation. Two of those approaches are certain combinations of feature selection and classification models. I must make a note of the obvious limitation of this review: I am not going to assess the general usefulness of certain classification models and feature selection algorithms in choice data. What is also obvious is that some peculiarities of various datasets may lead to situations favoring different methods. That said, all MD and conjoint data share some attributes which can lead to a preference of some methods over others. Therefore, the core questions in this part are: 1) Are all those methods useful for typing tool construction? 2) Should we prefer some of them? If so, always or only in some cases?

Below are the specifics of implementation of each approach:

### Standard and Adaptive Naïve Bayes

Standard—Johnson & Orme's (2009) greedy selection with Naïve Bayes classifier. 20 replicates to avoid local maxima.

Adaptive—based on above, 2 last item sets adaptive. Adaptive sets based on simple filter maximizing differences between probabilities for 2 most probable segments from preliminary exercises.

### Regression on Rankings

Full rankings as input. Multinomial regression as a building block. Feature selection: forward stepwise algorithm with exhaustive search in all possible additions on each step. Feature selection outside the cross-validation.

### Pairwise Classifier

Feature selection with recursive feature elimination based on boosted trees.

Three different classifiers: simple classification tree, random forest, linear discriminant analysis.

To provide a measure of classification efficiency, all the results were compared using average accuracy of classification across segments. To assure unbiased accuracy, I used 7-fold cross validation.[1]

Before the model building and testing process, raw utilities were disturbed with a Gumbel error of the form: -ln(-ln(X)) where X is uniform random variable. The main purpose of this step was to provide more real life accuracy assessment. If this step was omitted, the accuracy levels would be much too optimistic. All approaches would suffer from it in the same degree, although I have not put this assumption to the test.

## MaxDiff

The first and the most important type of data for our study should be MaxDiff. I decided to use three very different datasets to check the behavior of each strategy of building a typing tool:

[1] The main dataset was divided into 7 parts and then the whole model building was done leaving out one of the parts. This left-out part was used for testing accuracy.

- 20 items, 5 segment solution, N = 1081 which could be considered an average MaxDiff-based segmentation. The number of items is modest and segments are quite well separated.
- Extremely sparse data with 36 items and 5 segments, N = 700. Each of the items was shown only once during main study resulting in fuzzy borders between segments and potentially lots of uncertainty for classification.
- 4 very well separated segments on 12 items, N = 500.

To provide a more condensed view on the results, accuracy was averaged for all datasets. This decision was made after examining the results separately for the potential interactions between classifier accuracy and data conditions—none were found. The main difference stemmed from cluster separation and resulted with up or down shifts in accuracy for all classifiers.

## Results

The results presented below cover a comparison of typing tools built on pairs (Chart 1), triplets and quadruplets (Chart 2). The last two are examined only with Naïve Bayes and Regression on Rankings. For clarity, adaptive Naïve Bayes is compared only with standard Naïve Bayes (Chart 3).

**Chart 1**

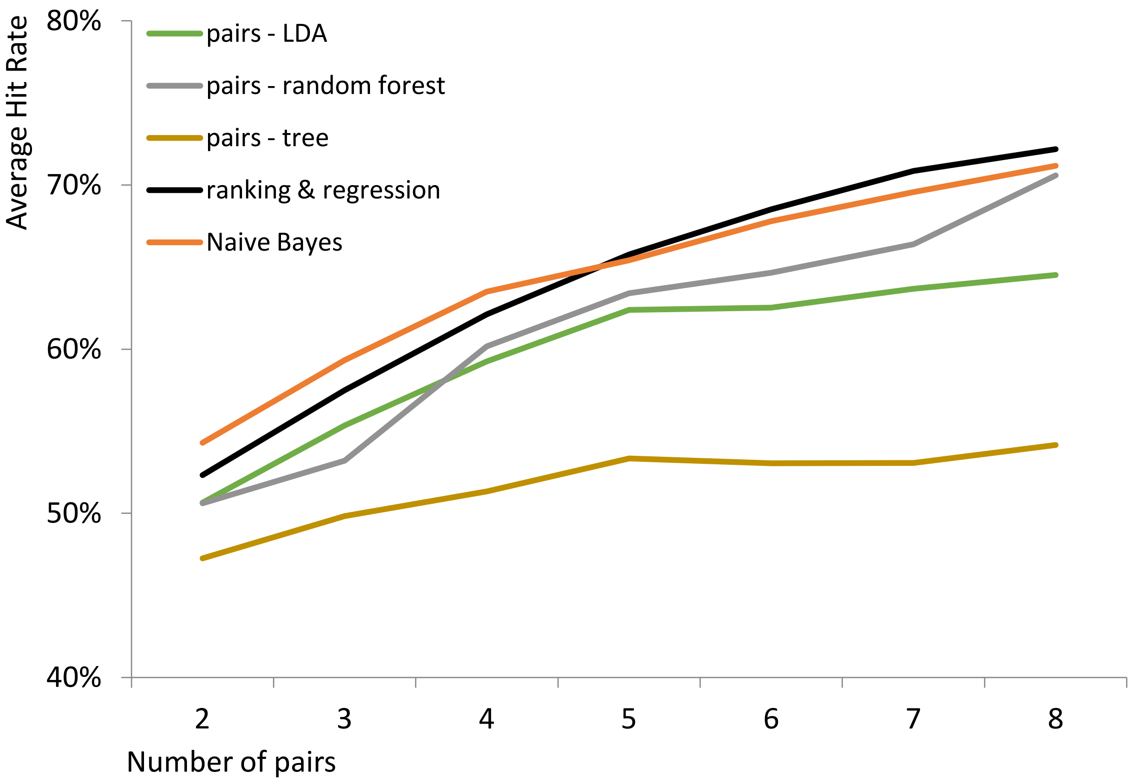
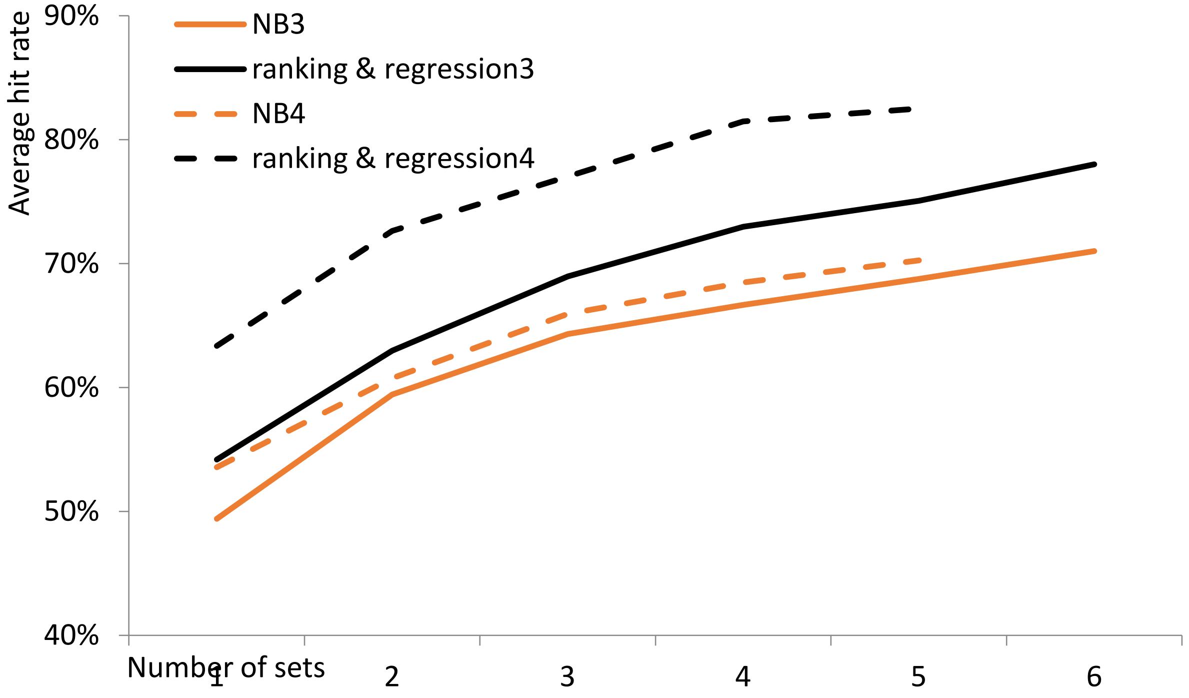


Chart 2



The results clearly demonstrate that starting from even as little as 2 pairs we get a large gain in classification power compared to the baseline of random guesses. We also boost accuracy with every additional question, but those gains diminish as we approach longer questionnaires.

Naïve Bayes and Regression on Rankings are best for any number of exercises explored and are very close to each other. The latter improves with the length of the questionnaire.

With pairwise classifiers we clearly depend on the classification algorithm. Random forest performs nearly as well as the winners, while a simple classification tree performs poorly (starting low and getting the smallest gains).

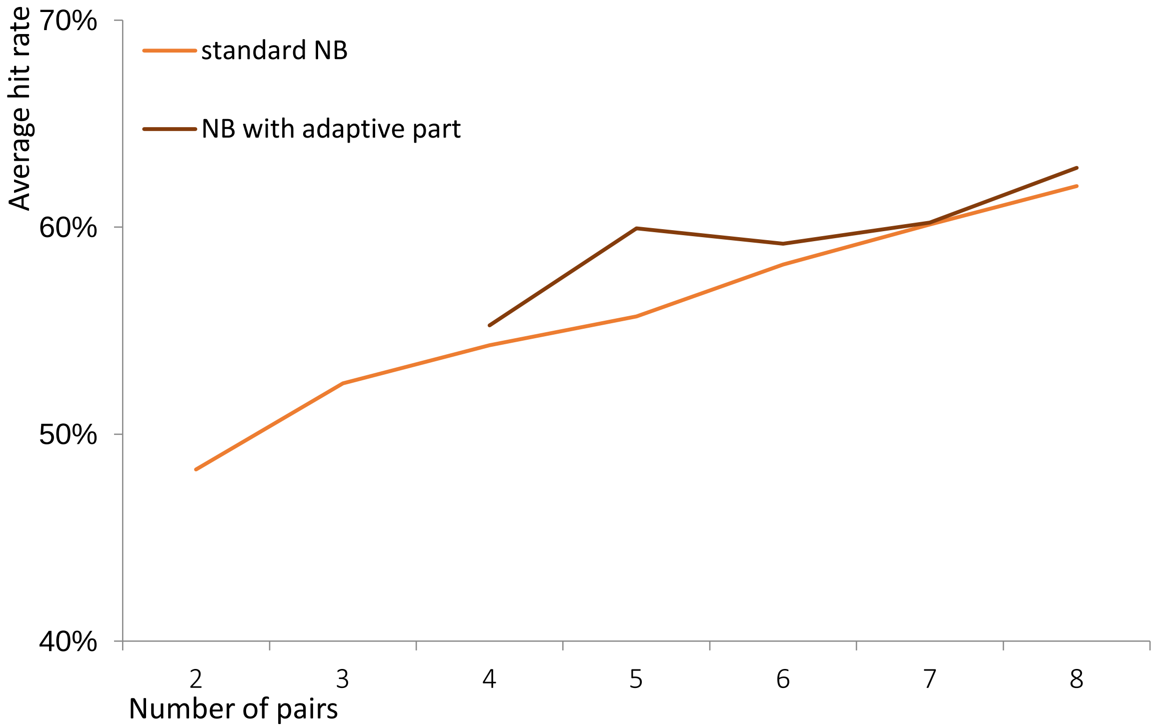
When we compare results for triads and quadruples exercises, the top position of Regression on Rankings is clear. Its gains are also much higher than in Naïve Bayes when more items are added to the set.

There are two caveats regarding the dominance of Regression on Rankings. First—this is the only tool using exhaustive search at each step of building a typing tool. For reasons of speed, feature selection was not included the cross-validation loop, so we can have legitimate concerns about possible biasing on the feature selection step. The second issue is that for quadruples we assumed a full ranking of all alternatives (compared to the best-worst exercise). In the case of live respondents, this could be much harder than B-W or a best-only question, and results in larger response error then the latter options. In this work, we assumed a uniform error for all situations which is, of course, a simplification—possibly biasing up results for larger sets and especially for quads with ranking.

## Probing Adaptive NB Classifier

Given the work involved in reprogramming the Naïve Bayes classifier and adaptive questionnaire, results of the adaptive part were quite disappointing (Chart 3). We only got 1% over standard NB. Such gain could be an interesting route for some machine learning competition, but in practice, it is immaterial. Of course, we could go into details of our implementation and improve some elements of this algorithm such as the criteria for optimal adaptive questions.

**Chart 3**

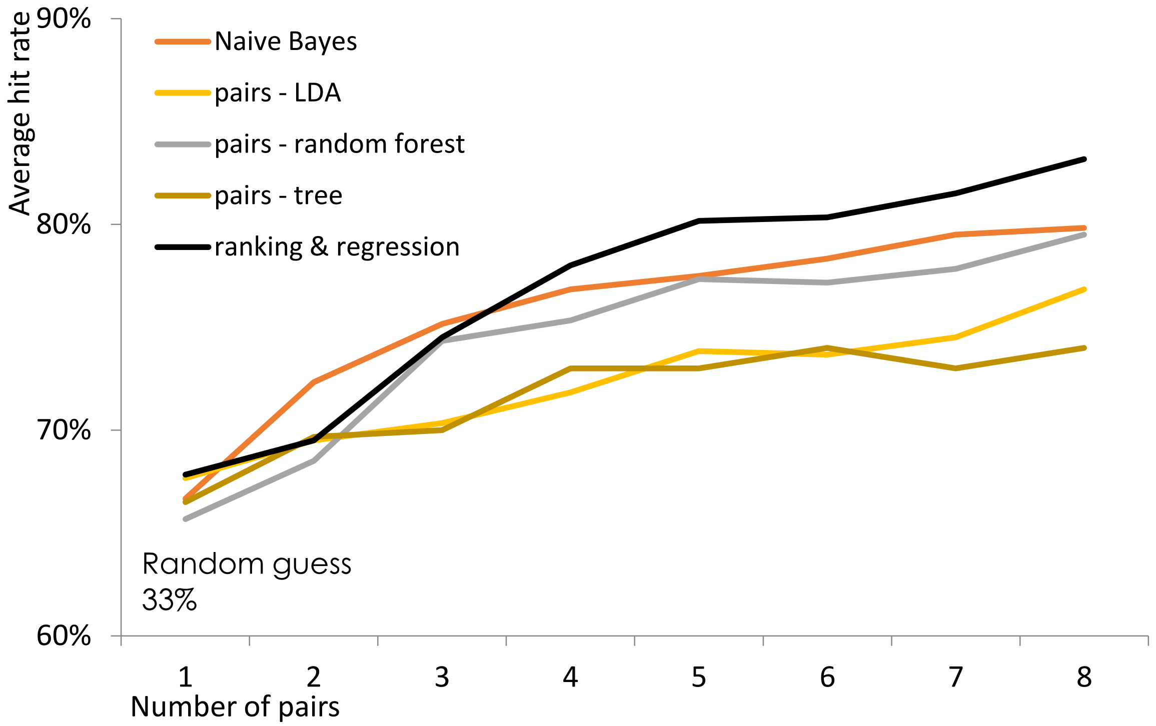


## Classifying CBC-Based Segmentation

The exploration of classification accuracy of typing tools derived by examined methods was done on one data set only, which came from a quite standard brand-price-pack study. The specifics were as follows:

- 4 attributes (4x4x4x2 levels), 3 segment solution, 600 respondents. The segmentation itself was done with CCEA on zero-centered HB utilities.

Since the main goal was to explore the adaptation of MaxDiff tailored algorithm to CBC data we experimented only with classifiers using pairs in discrete choice exercises, assuming that if it works with pairs it will work with larger datasets. That combined with a modest number of attributes and levels allowed us to follow the “all possible profiles” strategy for pairwise and regression algorithms. We got 128 profiles expanding into 8,128 pairs. As mentioned earlier, the Naïve Bayes algorithm added depth into the greedy search loop.



Strikingly, the entire pattern of results is very similar to one observed for pairwise classification MaxDiff data. Classification of this particular segmentation proved to be quite simple with only one pair more than doubling the random guess. Regression on Ranking won over all other approaches with more than 3 pairs. Also, interestingly, HB was the winner with shorter questionnaires. This last observation leads to the conclusion that the NB approach can have a tendency for local maxima which is more pronounced with more exercises. The greedy search algorithm most likely has a problem with finding the global maximum when the space of possible solutions is very large. Running more random starts or looping the search several times for each random start could probably improve its results.

## ACBC and Using BYO Exercises in Typing Tools

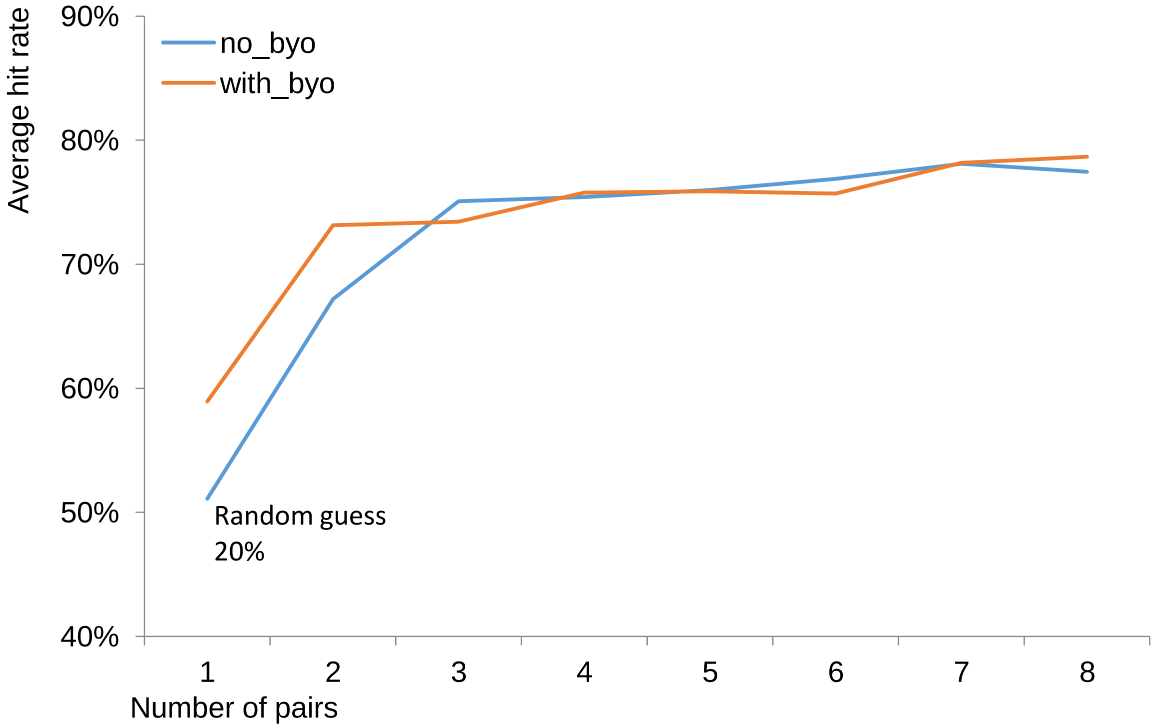
For ACBC-based segmentation we can use standard CBC-like choice questions (like those in the ACBC tournament). Elicitation of the optimal classifier will be conducted in the same way as for standard CBC. Of course, most ACBC problems are much more complex, so some prescreening of attributes for approaches other than NB is a must. Any search itself has to take a consideration of prohibitions, conditional relationships and summed price, which are quite common in ACBC studies.

What brings my exploration to the ACBC-based segmentation is a possibility of using a BYO question—which can be used in any typing tool but in ACBCs they are part of the estimation of utilities, so it seems more natural. A BYO question is generally liked by respondents, so it would be a bonus to break the routine of repetitive discrete choice questions and give them a more enjoyable task. As regards classification, it is expanded in the same way as for the standard estimation routine (to as many sets as there are attributes included in this question).

The study used here:

- 11 attributes with 6 variables in BYO part, 4 segments, 1040 respondents
  For simplicity, only NB with pairs was used. The BYO exercise was static and included all 6 variables used in the original study for this part.

**Chart 4**



It turns out (Chart 4) that the BYO exercise was of little value except for the shortest questionnaires. This result is disappointing, but we have to take one thing into consideration before we exclude the possibility of using BYO for typing tool building.

As mentioned above, the original BYO task had only ½ of all conjoint attributes. What is more, those included were not the most important discriminators of segments. We could hypothesize that the impact of BYO could be greater if BYO attributes were more important in segmentation.

## Conclusions

We found that each one of the explored options can work well for choice-based segmentation (MaxDiff-based, CBC-based alike). The "complete" approaches—joining specific feature selection and a classification algorithm worked best. Naïve Bayes with greedy search was tailored by its proposers (Orme & Johnson 2009) to work along the logic of MaxDiff. The greedy search can be seen as not optimal with more complicated problems—segmentations where the classifier must have many questions. However, it is the price to pay for excellent scalability—it works fast enough without any modifications for very large sets of items (and CBCs). This cannot be said about Regression on Rankings whose great results rely on stepwise exhaustive search.

On the other hand, logistic regression does not assume the features' (subsequent B-W answers) conditional independence. Sometimes the segmentation by a latent class will give us segments which comply with this assumption, but much more often ones that do not. In such cases Regression on Ranking will stand a considerable chance of providing a better model. This could also be one of the reasons why NB loses against regression with more questions in the typing tool. Of course, we could remove this assumption from NB but it would no longer be naïve and simple.

The pairwise classifier is more of an open framework than a ready solution. Given the right choice of the underlying classifier and algorithm for question sets it can produce excellent results. One of such choices is the pairing of a random forest with recursive feature elimination. But the main advantage of this approach is its simplicity—we can employ it with very basic analytical toolset. Just run some ranking of discrimination power of the items, build all the pairs, run a stepwise logistic regression or discriminant analysis and we arrive at a decent typing tool. In less than an hour and with no programming!

As to numbers, more is better, needless to say. We still have to consider the burden of longer questionnaires and/or more items in each exercise. But on average we can delete one exercise from a typing questionnaire per one added item in set. The gain is more pronounced when we have more exercises. Of course, each data set is different and accuracy increments will diminish at a different rate, so the only way to find our practical optimum is by experiment.

Lech Komendant

## Bibliography

Orme B., Johnson R. (2009)—A Procedure for Classifying New Respondents into Existing Segments Using Maximum Difference Scaling.

Pemberton J., Powlett J. (2006)—Identification of segments determined through non-scalar methodologies. Sawtooth Software Conference Proceedings.

# FULL-FLAVOURED HB: BYO DATA IN THE UPPER MODEL

**JANE TANG**
**ROSANNA MAU**
*MARU/MATCHBOX*
**MONA FOSS**
*BOOTSTRAP ANALYTICS*

## SUMMARY

Build-Your-Own (BYO) questions are quite common in CBC research. They are useful as a training/education tool prior to conjoint questions. Rather than throwing the BYO data away or encoding the responses into the choice data, BYO responses can be included as covariates in the upper-level HB model.

Covariates are most useful if they are related to attribute preferences. BYO questions on the conjoint attributes fit that requirement perfectly. The use of BYO as covariates adds nuanced, subtle, yet meaningful variation to the respondents' part-worth utilities; it captures greater heterogeneity and brings out the "full flavour" of HB.

Does the use of BYO questions as covariates improve the predictive validity of the models? That depends on the sparsity of the data and the amount of heterogeneity (disagreement across respondents). The sparser the data, the more likely prediction can be improved. At the same time, the more disagreement, the more opportunity there is for predictive gain by employing BYO questions as covariates.

The use of BYO questions in the upper-level HB model also provides a ready-made solution for generalization to future samples. The BYO questions themselves become the "golden" questions that can be quickly administered to new respondents—allowing researchers to apply the HB model to the new sample without the conjoint exercise.

## 1. INTRODUCTION

### 1.1 BYO

Build-Your-Own (BYO) exercises have long been a favourite tool for product development research. Product features are chosen one at a time, but these choices are made in the context of the full product combinations. The additional cost associated with advanced features can also be included as part of the exercise.

Sawtooth Software's Adaptive Choice-Based Conjoint (ACBC) uses BYO exercises as the basis for near neighbor configurations presented in the conjoint tasks. These data are then included as additional choice tasks in model estimation.

BYO exercises can also be used for respondent education. In earlier research, Tang *et al.* (2009) showed that when they are used as an introduction to a discrete choice exercise, BYO questions are useful in familiarizing respondents with the product features being tested and focusing each respondent's attention on his most salient features. They can also potentially impact derived price sensitivity in the subsequent conjoint exercise.

## 1.2 The Upper-Level Model

While useful for educating respondents, the BYO data itself is typically not used for anything. The question is, could it be? One option would be to code the BYO answers as additional choice tasks, as in ACBC. Alternatively, this research set out to determine if BYO data could be used effectively in the upper-level HB model.

The upper model is what makes Hierarchical Bayes *hierarchical*. The lower model is what most researchers are familiar with, that is, the individual-level utilities, or the beta estimates. The means of these utilities are determined by a multivariate regression model—the "upper model."

In the simplest case, one with no covariates, a single intercept, is used to model the mean. This is an "uninformative" prior, one that assumes all respondents have similar preferences. Adding covariates to the upper model enables it to estimate different upper-level means given the covariate pattern.

For example, if we include gender in the upper model, we are making the assumption that males and females have different preferences, that the means of the betas are different for each gender. These differences will appear in the alpha. To use an analogy, an HB model without covariates is an everyday meal. It is good and nutritious. However, if you want a really tasty meal, you will want something more elaborate. The covariates in the upper model are what give HB its full flavour.

## 1.3 Previous Research

It is important to consider which variables should be used as covariates. Upper-level variables can have an impact on lower-level parameters only if they are strongly correlated with the attribute preferences or purchasing behaviors being modeled. They are also more impactful with sparse data (Eagle 2016). However, it is not a good idea to throw everything into the upper model in the hopes of finding something useful. Spurious covariates can hurt the model. Badly fitting covariates can adversely affect lower-level parameters and overfitting can influence results.

Other studies have looked at the use of covariates in the upper model. McCullough (2014), Sentis & Geller (2010), Kurz & Binner (2010), all produced similar results. McCullough summarized them succinctly, saying the purpose of covariates in the upper level of the HB model is probably *not* to improve model performance, as measured by hit rates, MAE, etc. The purpose probably *is* to better describe or understand respondent heterogeneity.

It is important to note that in all three studies referenced here, potential covariates were used one at a time. McCullough (2014) was the only one to use BYO data—brand and price—but again, one variable at a time. Sentis & Geller (2010) used in-sample statistics to compare model performance based on holdout tasks. Kurz & Binner (2010) used both in-sample statistics and real market data. McCullough (2014) based his model performance comparisons on holdout sample, including hit rates.

## 1.4 Prediction in Holdout Sample

Evaluating model performance using the holdout sample is generally preferred over that of using "in-sample" holdout tasks (Eagle 2016). However, there are many different ways one can generalize from a Hierarchical Bayes model (Pachali *et al.* 2014). Perhaps the more theoretically "correct" way is through the HB draws. However, for most practitioners, that's often difficult to do. For our purposes, we generalized the results of our HB models through the posterior means. That is, we used the posterior alpha draws, calculated the average across the draws to create posterior means. We then run the holdout sample data through these means to create a pseudo "beta" like utility estimate for each respondent in the holdout sample. We use these pseudo "beta" values to make predictions for each respondent in the holdout sample.

## 2. Objectives, Data & Results

The purpose of this research was to determine the impact of BYO data specifically on respondent heterogeneity and on hit rates. We wanted to know, can BYO data bring out flavour? To this end, we compared models with BYO covariates to models with no covariates, as well as to those with other more traditional covariates, such as demographic data.

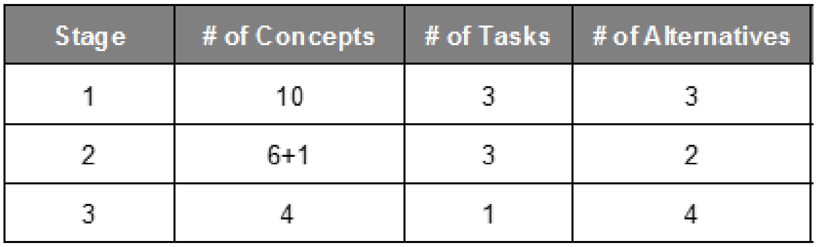
We also wanted to know if the number of choice tasks is important. Does it make a difference if we have sparse data?

And finally, how should we prioritize when we have too much BYO information realistically to run an HB model quickly. Is it better to pick the variables that matter, or to use segments derived from the BYO data?

To answer these questions, we looked at three real world studies where BYO data was collected as part of an education section prior to the conjoint exercise: 1. a MaxDiff concept test; 2. a CBC study about Congressional politics and; 3. a CBC study about women's dating preferences.

### 2.1 Study 1: MaxDiff Concept Test

The first study we looked at was an Adaptive MaxDiff exercise involving 10 concepts, with binary anchoring to none to indicate purchase. An adaptive MaxDiff exercise is simply a MaxDiff exercise repeated in stages. Items chosen as worst are dropped off in each stage, and the later stages focus on the more preferred items.



| Stage | # of Concepts | # of Tasks | # of Alternatives |
|---|---|---|---|
| 1 | 10 | 3 | 3 |
| 2 | 6+1 | 3 | 2 |
| 3 | 4 | 1 | 4 |

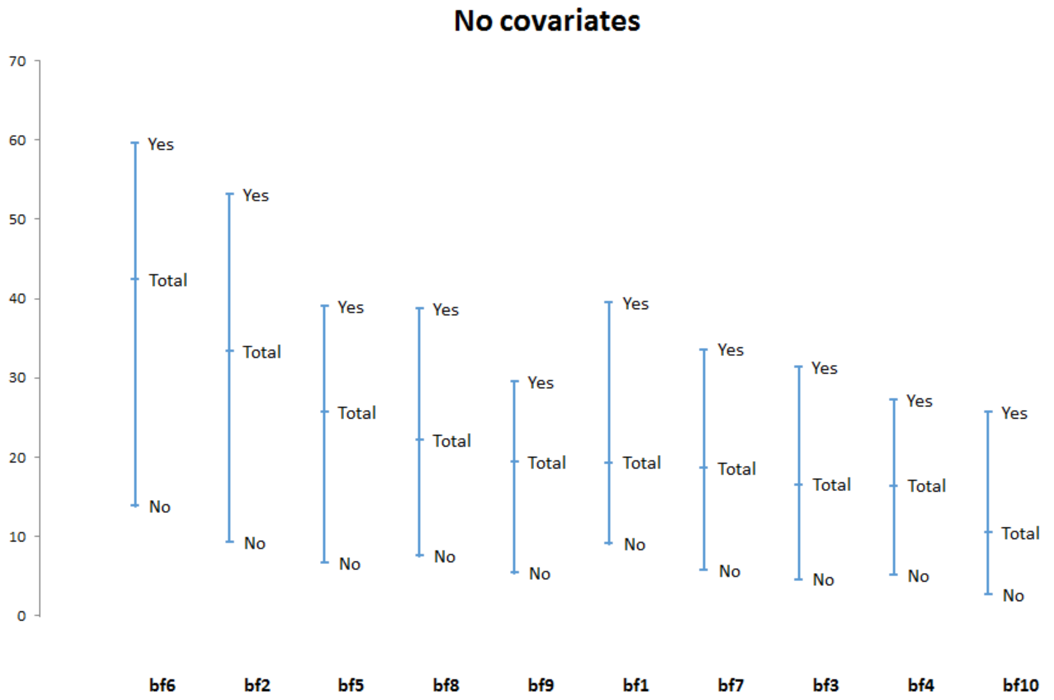
For example, here in stage one, nine of the ten concepts were randomized to three tasks, each with three options. Respondents were asked to choose the one they preferred the most and the one they preferred the least in each task. The least preferred concept from each task was discarded. In stage two, the remaining six concepts, plus the one not included in stage one (total seven concepts) were shown as three paired comparisons. The one concept not shown in the three pairs in stage two was carried forward to stage three. The three concepts not chosen as preferred in the paired comparisons were dropped. In stage three, the four remaining concepts (three preferred concepts from stage two and one concept carried forward) were ranked based on the respondent's order of preference.

Respondents were introduced to the ten concepts in an education section prior to completing the MaxDiff exercise. As in many concept tests, each concept was given a descriptive name to help identify it in the MaxDiff exercise. However, it was important to be sure respondents had read and understood what each concept was about before completing the MaxDiff tasks. This was accomplished by showing the concepts one at a time and asking respondents whether or not they were interested in each one.

From this simple exercise we obtained a set of binary variables—interested or not—for the ten concepts. Normally this data would not be used for anything—it was simply a means to engage respondents in the concept descriptions, to force them to read and think about them. However, we wondered, could this throw-away data be useful? Could it be used to improve the model?

To answer this question, we first ran a simple model with no covariates and plotted the simulated purchase intent for each concept. In Figure 1-1 below, the “Total” on each line indicates the mean purchase intent. The “Yes” and “No” points show the means for two groups—those who expressed interest in the concept in the education section, and those who did not.

**Figure 1-1**



A casual comment from the client when he saw this chart was, “Oh, I was expecting to see greater differences between the groups, based on their interest.” And while there were differences, they were not as striking as the client expected. Could this be because we assumed in the upper model that all the respondents come from a single, uninformative prior?

Another way to see the differences between these two groups is to look at the distributions of the betas. Figure 1-2 below shows the distribution of the beta for one of the concepts, Concept D, grouped by initial interest in it. While there is good separation between the two groups, F=259 with (1, 1500) degrees of freedom, there is also a large overlap.

**Figure 1-2**

**No Covariates**

| | F stat |
|---|---|
| No Covariates | 259 |

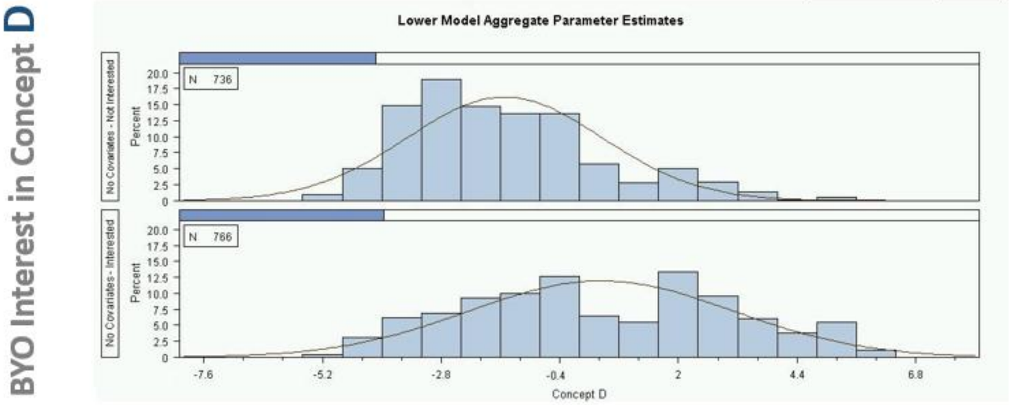
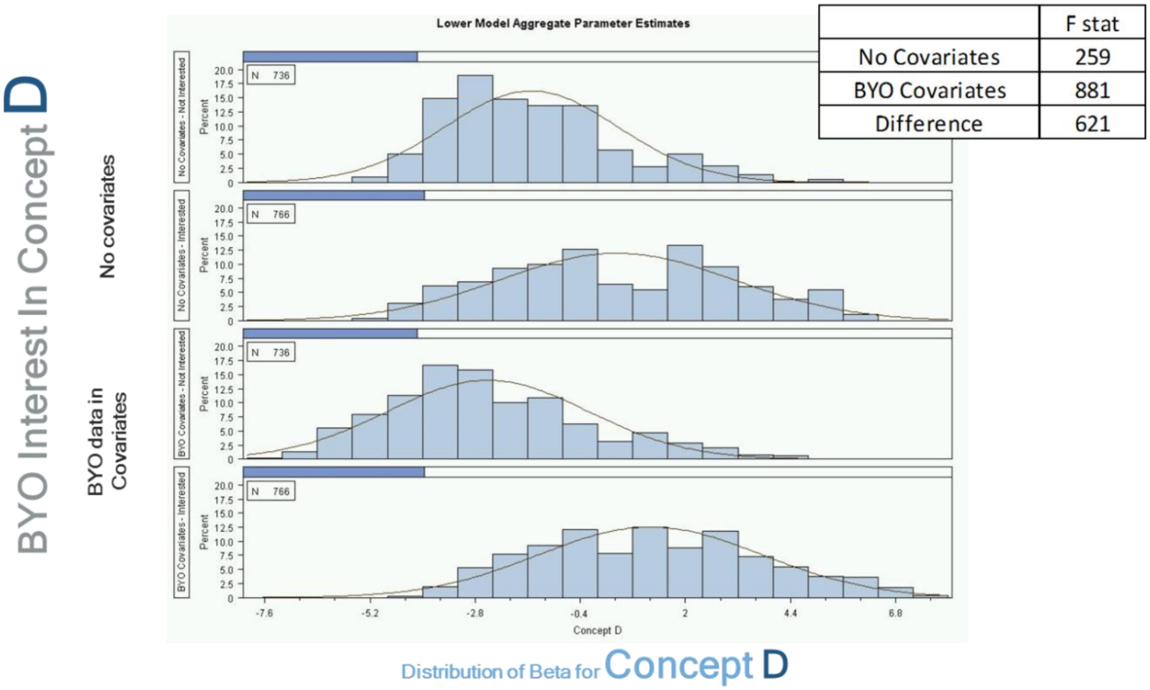


Figure 1-3 compares this simple model where we have no covariates to one that includes all ten BYO variables in the upper model. Again, looking at the distributions of the beta for Concept D, the impact is quite clear: BYO covariates allow us to more effectively separate those who are interested in Concept D, versus those who are not. The F statistics also support this conclusion, increasing from 259 in the model with no covariates to 881 in the model with covariates, both with (1,1500) degrees of freedom.

**Figure 1-3**



| | F stat |
|---|---|
| No Covariates | 259 |
| BYO Covariates | 881 |
| Difference | 621 |

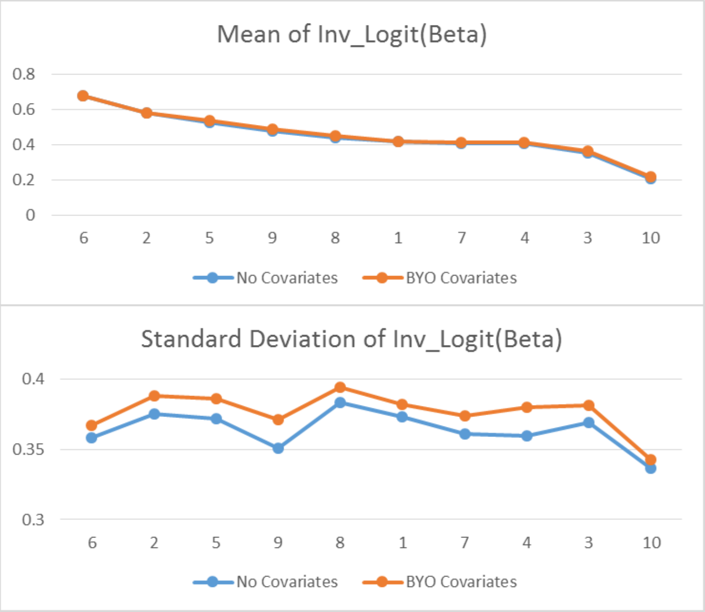
We see this pattern for all ten concepts in the table of differences in F-statistics in Figure 1-7. The result of including the BYO data in the upper model is more effective separation. The separation is meaningful—it is where you expect it to be—and it leads to higher overall variance of the betas.

Another way to show this increase in overall variance in the model with BYO covariates is in Figure 1-4, which plots the summary statistics for the inverse logit of the beta for each concept, and for each model. The inverse logit transformation was used to deal with the scale factor issue that is inherent in the logit estimate.

$$f(\beta) = \frac{e^{\beta}}{1+e^{\beta}}.$$

It is well known that the raw beta utility score from the HB process is tied to a scale parameter; the better the model fit, the larger the magnitude of the estimate. The transformed betas are much easier to work with, having a range of between 0 and 1.

**Figure 1-4**

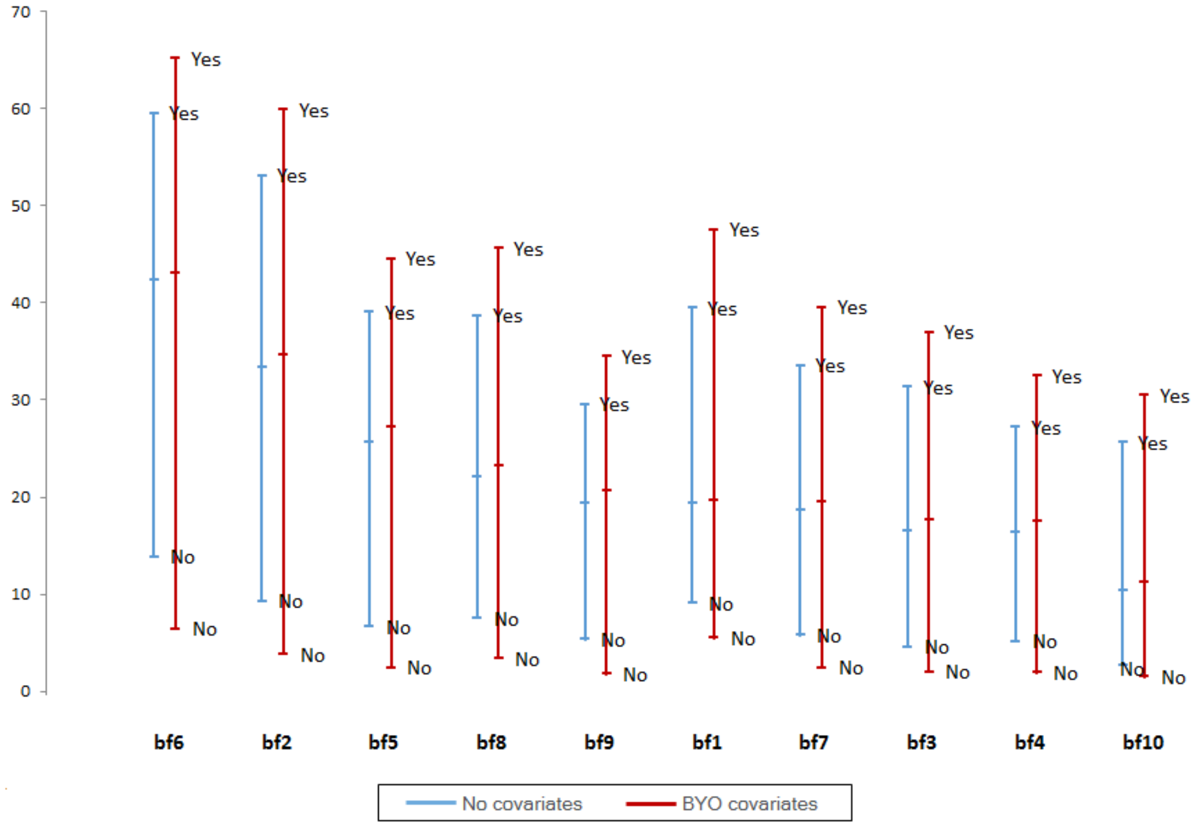


The top half of the chart clearly shows that the means of the betas do not change when BYO covariates are included. However, as a result of better separation in the distributions of the betas, there is an increase in the variability across individual respondent's beta estimates, which is evident in the standard deviation plot in the lower half of Figure 1-4. This increase in variance is a reflection of the heightened ability to capture individual heterogeneity using BYO covariates, compared to no covariates.

While we will use this increase in variance of the betas as an indicator for meaningful separation throughout this paper, the best way to check for it is to plot out the distributions of the betas. Greater variance is not enough in itself—there needs to be real separation, not just greater variability—and the separation needs to be meaningful.

Now we come back to our client, and his initial comment about the differences in mean purchase intent for the concepts, based on interest, not being as great as he expected. When we add the BYO data to the upper model we see a much more striking picture than what we saw initially (Figure 1-5).

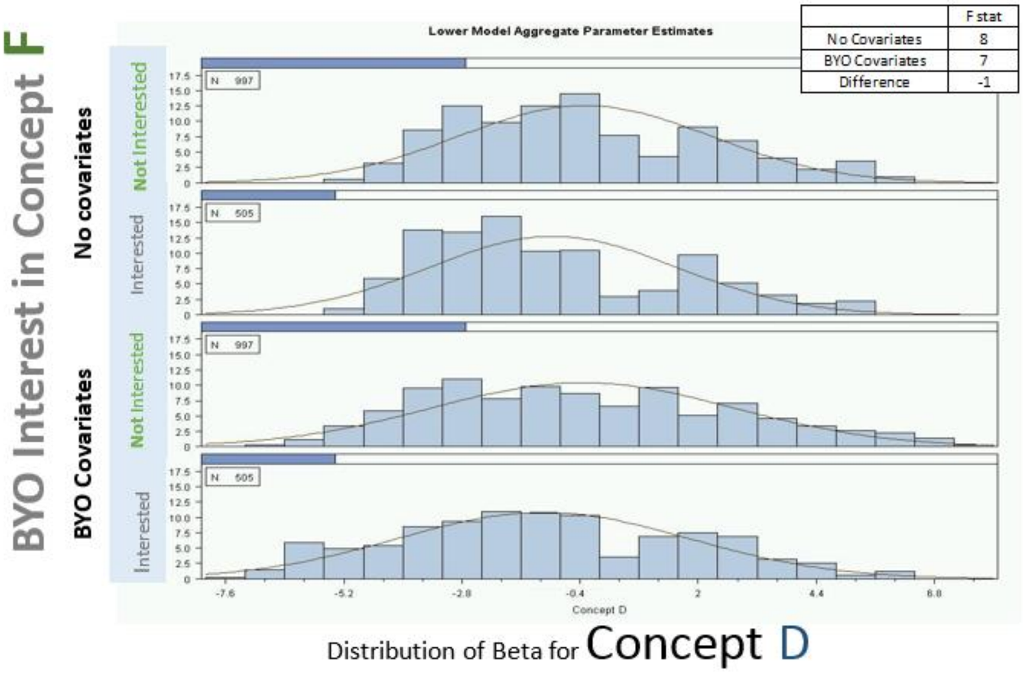
**Figure 1-5**



While the mean for each concept remains the same as it was in the initial model with no covariates, we now see greater contrast between respondents who are inherently interested in each concept, versus those who are not. Using the BYO data in the upper model adds nuance—subtle, yet meaningful variation in the betas. The results are more palatable for the client, and easier for him to understand. It adds flavour to our HB model.

One of the differences between this work and previous research into upper model covariates is that here we have introduced several variables into the upper model simultaneously. One concern with this approach is the possibility of overfitting. Sentis & Geller (2010) demonstrated evidence of overfitting in the increased separation between analysis groups by introducing a randomly scrambled segment variable in the upper model. Along this same line of thinking, we looked at whether initial interest in an unrelated concept, Concept F, would impact the utility for Concept D. Separation in the distribution of the beta for concept D for those interested/not interested in Concept F would be evidence of overfitting; it would indicate separation that is not meaningful. However, if there is no overfitting, the distributions should overlap completely.

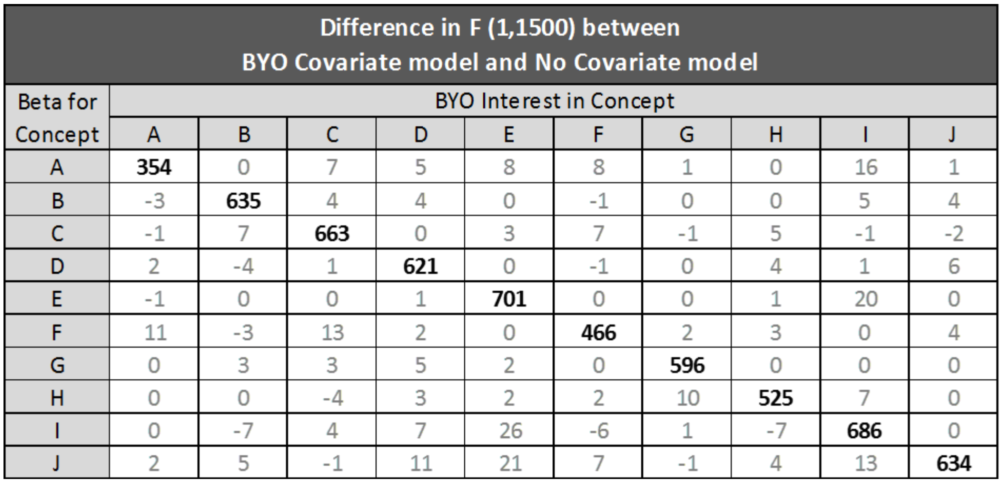
**Figure 1-6**



| | F stat |
|---|---|
| No Covariates | 8 |
| BYO Covariates | 7 |
| Difference | -1 |

Figure 1-6 shows the distribution of the Concept D beta as before, but this time grouped based on interest in Concept F. There is no separation between those interested in Concept F and those not interested. The distributions overlap completely in the model without covariates and also in the model with BYO covariates. This is as it should be, since we would not expect interest in an unrelated concept to impact preference for Concept D. Additionally, the F statistics are very small, again with (1,1500) degrees of freedom, and do not effectively change with the addition of covariates. There is no evidence of overfitting despite the inclusion of the ten BYO covariates.

**Figure 1-7**



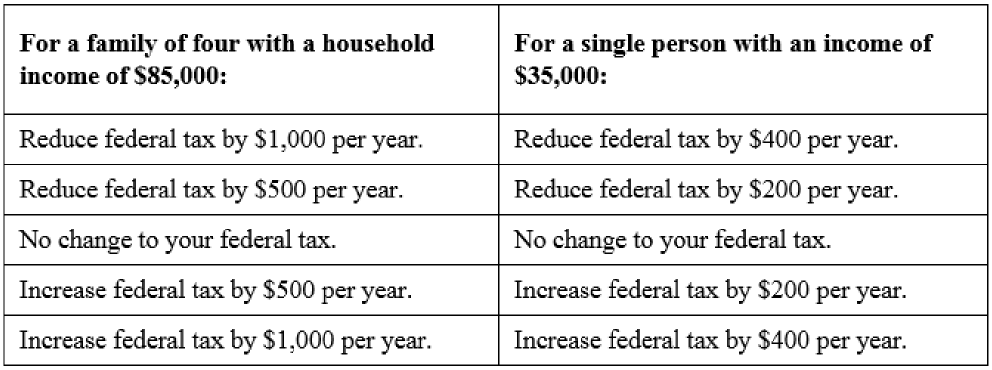
| Difference in F (1,1500) between<br>BYO Covariate model and No Covariate model | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Beta for | BYO Interest in Concept | | | | | | | | | |
| Concept | A | B | C | D | E | F | G | H | I | J |
| A | **354** | 0 | 7 | 5 | 8 | 8 | 1 | 0 | 16 | 1 |
| B | -3 | **635** | 4 | 4 | 0 | -1 | 0 | 0 | 5 | 4 |
| C | -1 | 7 | **663** | 0 | 3 | 7 | -1 | 5 | -1 | -2 |
| D | 2 | -4 | 1 | **621** | 0 | -1 | 0 | 4 | 1 | 6 |
| E | -1 | 0 | 0 | 1 | **701** | 0 | 0 | 1 | 20 | 0 |
| F | 11 | -3 | 13 | 2 | 0 | **466** | 2 | 3 | 0 | 4 |
| G | 0 | 3 | 3 | 5 | 2 | 0 | **596** | 0 | 0 | 0 |
| H | 0 | 0 | -4 | 3 | 2 | 2 | 10 | **525** | 7 | 0 |
| I | 0 | -7 | 4 | 7 | 26 | -6 | 1 | -7 | **686** | 0 |
| J | 2 | 5 | -1 | 11 | 21 | 7 | -1 | 4 | 13 | **634** |

We found the same pattern for all the concepts in Figure 1-7. Separation occurs only where it should, and is meaningful.

## 2.2 Study 2: Congressional Politics

Our second study is a dataset used in Tang & Grenville (2010). It used Choice-Based Conjoint to understand how respondents vote based on political policy platforms. Fieldwork was conducted in 2010 and included a BYO section as an introduction to the CBC, asking respondents to describe their ideal candidate's policy platform.

The CBC design included factors for five public policy areas: health care, foreign affairs, size of government, environment and education. Each had two levels describing the traditional Democratic and Republican party positions, as well as a third "no mention" level. A sixth factor described the federal tax implications of the policy platform, as below:



| **For a family of four with a household income of $85,000:** | **For a single person with an income of $35,000:** |
|---|---|
| Reduce federal tax by $1,000 per year. | Reduce federal tax by $400 per year. |
| Reduce federal tax by $500 per year. | Reduce federal tax by $200 per year. |
| No change to your federal tax. | No change to your federal tax. |
| Increase federal tax by $500 per year. | Increase federal tax by $200 per year. |
| Increase federal tax by $1,000 per year. | Increase federal tax by $400 per year. |

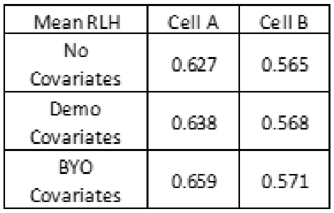
The BYO questions were based on the five policy areas only. The questionnaire also included a holdout task. Other data collected in the survey included the most important issue facing the US today, performance of the President, Congressional approval, voting intent, past (2008) vote, party affiliation, and political leaning (conservative vs. liberal).

The sample was divided into four cells. Cell A1 had 500 respondents, and each saw six choice tasks with three options. Cell B had a similar number of respondents (n=504) and 15 choice tasks, each with three options. A1 and B were used as calibration samples. Cell A2 had 712 respondents, and was set up the same way as Cell A1. Cell C had 308 respondents, and each saw 15 tasks with five options. A2 and C were used as holdout samples.

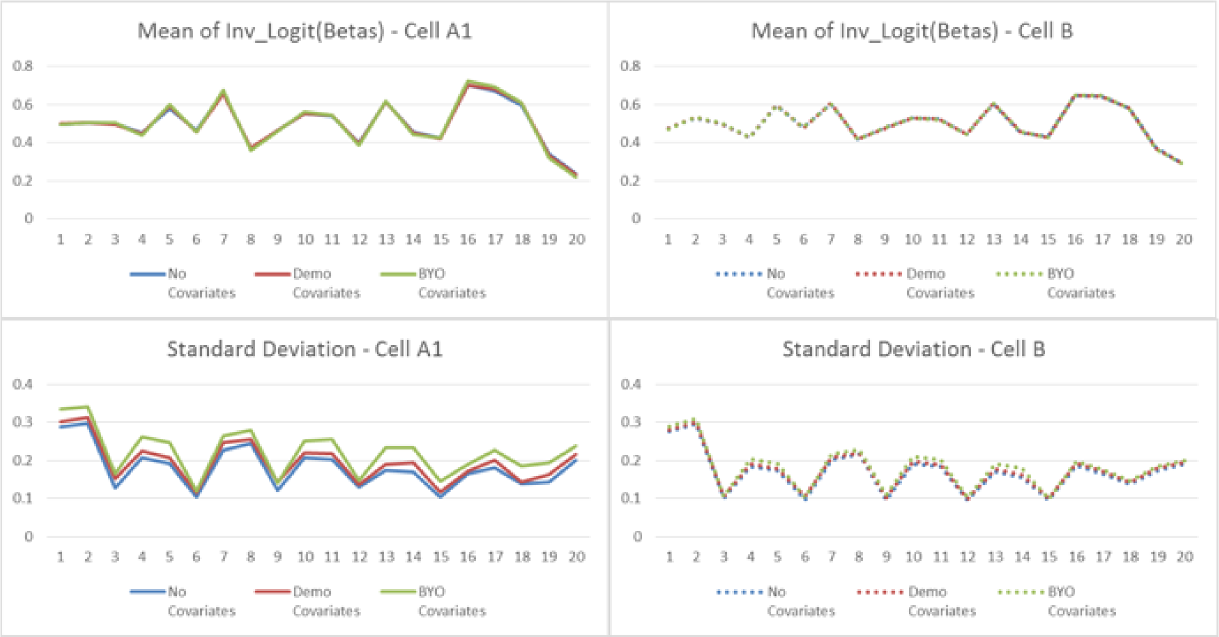
We ran three versions of HB models for each of Cell A1 and B: one with no covariates; one with "demographic" variables—political leaning and Congressional approval—as covariates; and one with the BYO data—five public policy areas—as covariates.

As we did with the MaxDiff example, we plotted the summary statistics for the inverse logit of the betas for each cell and each model, as shown in Figure 2-1.

**Figure 2-1**



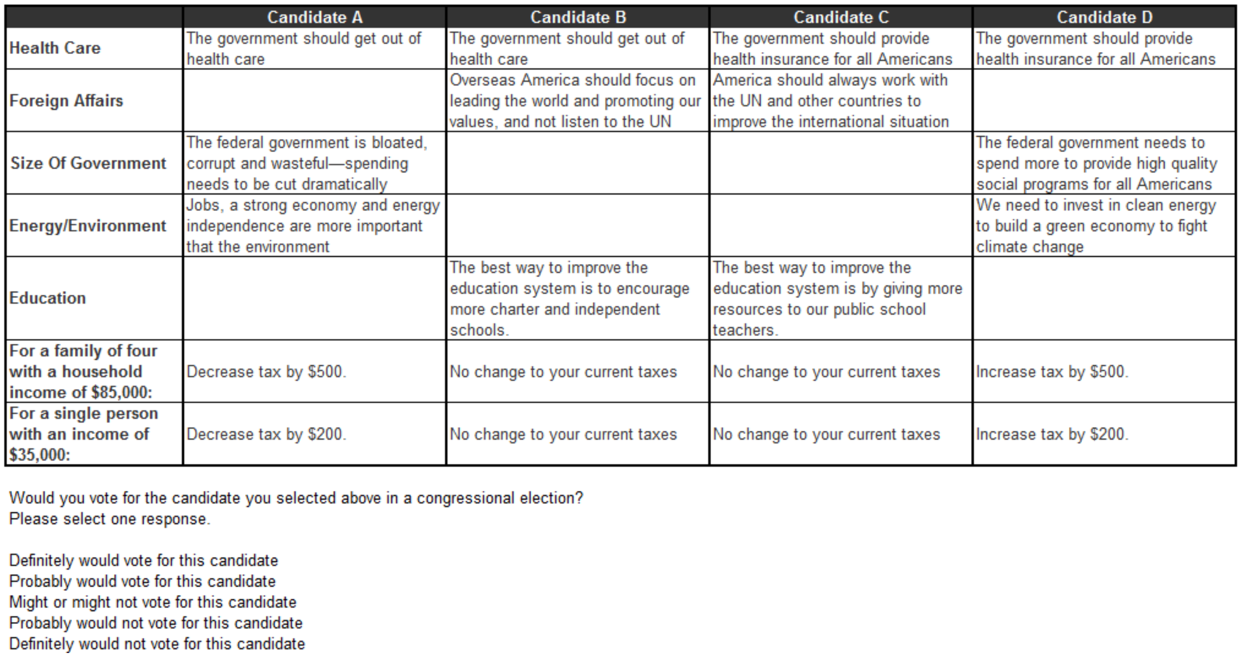
| Mean RLH | Cell A | Cell B |
|---|---|---|
| No Covariates | 0.627 | 0.565 |
| Demo Covariates | 0.638 | 0.568 |
| BYO Covariates | 0.659 | 0.571 |



Again, the covariates did not affect the mean of the betas, but they did help to improve the variability among respondents for Cell A1, where the data is sparse. The impact on Cell B, where there were more choice tasks, was considerably less.

As before, we are using the increase in variability as a shorthand indicator for “meaningful separation.” Adding covariates does not automatically increase variability/RLH/scale factor, as we see in Cell B. Covariates are only useful where it matters, that is, where the data is sparse and the covariates are meaningful. Notably, in Cell A1, where the impact of covariates is greater, the BYO covariates had a greater impact on variability than the demographic covariates.

We designed a holdout task with balanced alternatives for this study that was identical for all cells. It included traditional Democratic and Republican party positions, both moderate and more extreme. Respondents were asked to choose their preferred candidate out of the four options presented, and then asked how likely they would be to vote for that candidate in a Congressional election.



| | Candidate A | Candidate B | Candidate C | Candidate D |
|---|---|---|---|---|
| **Health Care** | The government should get out of health care | The government should get out of health care | The government should provide health insurance for all Americans | The government should provide health insurance for all Americans |
| **Foreign Affairs** | | Overseas America should focus on leading the world and promoting our values, and not listen to the UN | America should always work with the UN and other countries to improve the international situation | |
| **Size Of Government** | The federal government is bloated, corrupt and wasteful—spending needs to be cut dramatically | | | The federal government needs to spend more to provide high quality social programs for all Americans |
| **Energy/Environment** | Jobs, a strong economy and energy independence are more important that the environment | | | We need to invest in clean energy to build a green economy to fight climate change |
| **Education** | | The best way to improve the education system is to encourage more charter and independent schools. | The best way to improve the education system is by giving more resources to our public school teachers. | |
| **For a family of four with a household income of $85,000:** | Decrease tax by $500. | No change to your current taxes | No change to your current taxes | Increase tax by $500. |
| **For a single person with an income of $35,000:** | Decrease tax by $200. | No change to your current taxes | No change to your current taxes | Increase tax by $200. |

Would you vote for the candidate you selected above in a congressional election?
Please select one response.

Definitely would vote for this candidate
Probably would vote for this candidate
Might or might not vote for this candidate
Probably would not vote for this candidate
Definitely would not vote for this candidate

Comparing in-sample hit rates on the holdout task for each model and each cell, we saw slight improvement for Cell A1, where the data was sparse, but little or no improvement for Cell B, where we had more choice tasks.

| Hit Rate Holdout Task (Dual Response) In Sample | | | | |
|---|---|---|---|---|
| n=~500 | | No Covariates | Demo Covariates | BYO Covariates |
| Preference Only | Cell A1 (6 choice tasks) | 52% | 53% | 56% |
| | Cell B (15 choice tasks) | 57% | 58% | 58% |
| Preference + "No buy" | Cell A1 (6 choice tasks) | 54% | 55% | 57% |
| | Cell B (15 choice tasks) | 54% | 54% | 55% |

However, when we looked at the results for holdout sample (Cell A2 and C), the improvement in holdout task hit rates were more dramatic. Here the addition of demographic covariates improved hit rates substantially. And BYO data offered even more improvement.

| Hit Rate Holdout Task (Dual Response) Holdout Sample | | | | |
|---|---|---|---|---|
| Upper model developed for Cell A1 (In-sample data) | | No Covariates | Demo Covariates | BYO Covariates |
| Preference Only | Cell A2 Holdout Sample | 37% | 48% | 55% |
| | Cell C | 34% | 41% | 51% |
| Preference + "No buy" | Cell A2 Holdout Sample | 32% | 39% | 43% |
| | Cell C | 36% | 40% | 46% |

| Hit Rate Holdout Task (Dual Response) Holdout Sample | | | | |
|---|---|---|---|---|
| Upper model developed for Cell B | | No Covariates | Demo Covariates | BYO Covariates |
| Preference Only | Cell A2 Holdout Sample | 34% | 45% | 55% |
| | Cell C | 33% | 40% | 50% |
| Preference + "No buy" | Cell A2 Holdout Sample | 31% | 37% | 42% |
| | Cell C | 35% | 38% | 42% |

It is also interesting to note that the model developed using less data (Cell A1, 6 tasks) did slightly better than the model developed using more data (Cell B, 15 tasks) when they were generalized to the holdout sample.

However, the hit rate in holdout sample was still less than that of the in-sample hit rates for the constant holdout task, suggesting that despite the improvement brought on by the upper model, there was still some way to go compared to the in-sample prediction due to the contribution of the lower model itself.

## 2.3 Study 3: Dating

The dating study used in Tang & Grenville (2013), was fielded in 2013, to approximately 300 respondents in each of Canada and the United States, and looked at women's dating preferences. There were nine factors in the design, and nine corresponding BYO questions, shown below.

| Attribute: | Level: | Level: | Level: | Level: | Level: | Notes: |
|---|---|---|---|---|---|---|
| Age | Much older than me, | A bit older than me | About the same age | A bit younger than me | Much younger than me | |
| Height | Much taller than me | A little taller than me | Same height as me | Shorter than me | | |
| Body Type | Big & Cuddly | Big & Muscly | Athletic & Sporty | Lean & Fit | | |
| | | | | | | images used at the BYO question only, not in conjoint task |
| Career | Driven to succeed and make money | Works hard, but with a good work/life balance | Has a job, but it's only to pay the bills | Prefers to find work when he needs it | | |
| Activity | Exercise fanatic | Active, but doesn't overdo it | Prefers day to day life over exercise | | | |
| Attitude towards Family/Kids | Happy as a couple | Wants a few kids | Wants a large family | | | |
| Personality | Reliable & Practical | Funny & Playful | Sensitive & Empathetic | Serious & Determined | Passionate & Spontaneous | |
| Flower Scale | Flowers, even when you are not expecting | Flowers for the important occasions | Flowers only when he's saying sorry | "What are flowers?" | | |
| Yearly Income | Pretty low | Low middle | Middle | High middle | Really high | |
| | Under $50,000 | $50,000 to $79,999 | $80,000 to $119,999 | $120,000 to $159,999 | $160,000 or more | Australia |
| | Under $30,000 | $30,000 to $49,999 | $50,000 to $99,999 | $100,000 to $149,999 | $150,000 or more | US/Canada |
| | Under £15,000 | £15,000 – £39,999 | £40,000 – £59,999 | £60,000 – £99,999 | £100,000 or more | UK |

Respondents completed eight dual response CBC tasks, each with three alternatives. The research showed that women in the two countries are largely similar in their preferences, with US women slightly more likely to date their preferred alternative.

As there was no designed holdout task for this study, we randomly selected one of the eight choice tasks from each respondent to use as the holdout task. Additionally, since the two countries are largely similar, and arguably can be treated as being in the same market, we also used the US data as holdout sample to demonstrate the idea of generalizing the HB results to the market.

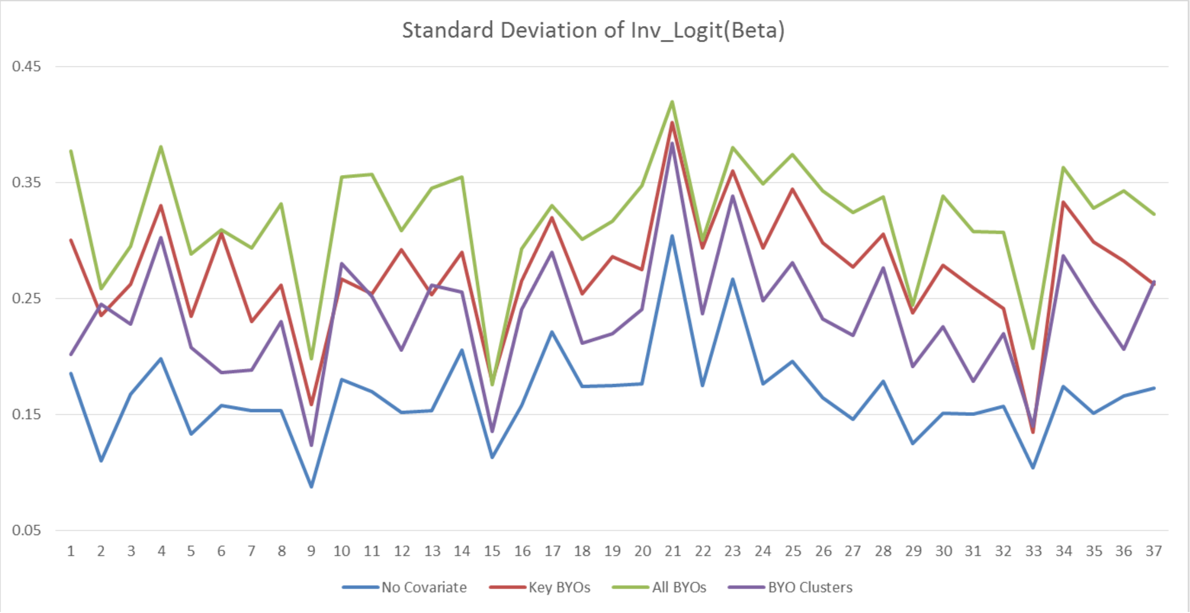
We ran four versions of the model, with widely differing times to complete the HB runs. The first model, with no covariates, took seven minutes. The second model, using all nine BYO questions as covariates, took 58 hours. At this point it became evident that we needed a more efficient way to include this much BYO data in the upper model.

The first alternative was to use six key BYO variables as covariates. These key questions were where we saw the most differentiation across the levels. They were selected using a simple counting analysis showing how often each level was chosen in each factor, and identifying where the largest differences appeared between the most and the least preferred levels. This model took six hours to run—quite a bit better than 58 hours, but still a fairly long run time.

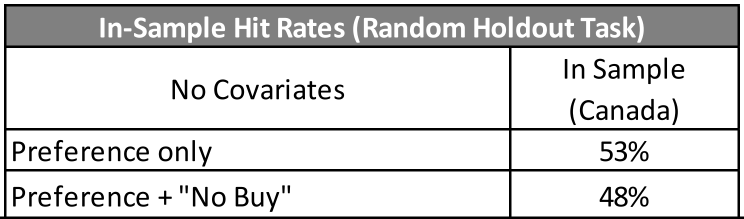
As a final alternative, we ran a latent class segmentation using all the BYO questions. Twelve clusters were derived and used as covariates in the upper model. This version took 65 minutes to run.

Figure 3-1 compares the standard deviation of the inverse logit of the betas for each of the four models. Again we see increased variance for the models with BYO covariates, compared to no covariates. The model with all the BYO covariates has the most variability, and the model with the BYO clusters has the least of the three BYO versions. However, as discussed, an important trade-off must be made between “flavour” (variability) and efficiency (run time).

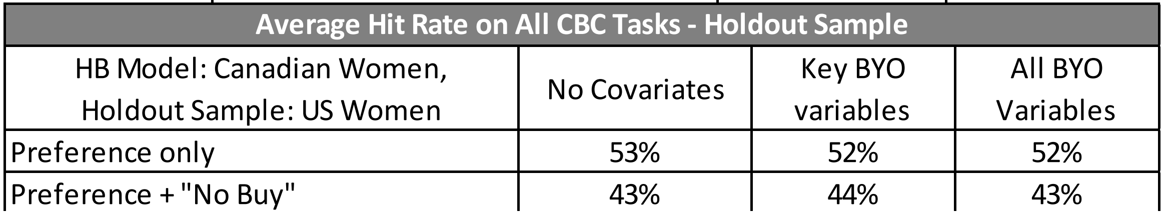
**Figure 3-1**



Turning to model prediction, we see a different pattern from the Congressional politics study. We found virtually no improvement in hit rates in the models where BYO data is used. However, the hit rates for the holdout sample are comparable to that for the in-sample rates.



| In-Sample Hit Rates (Random Holdout Task) | |
|---|---|
| No Covariates | In Sample (Canada) |
| Preference only | 53% |
| Preference + "No Buy" | 48% |



| Average Hit Rate on All CBC Tasks - Holdout Sample | | | |
|---|---|---|---|
| HB Model: Canadian Women, Holdout Sample: US Women | No Covariates | Key BYO variables | All BYO Variables |
| Preference only | 53% | 52% | 52% |
| Preference + "No Buy" | 43% | 44% | 43% |

Since there was no purposefully designed holdout task in this study, one of the eight CBC tasks was randomly selected in the Canadian sample to act as the holdout task in order to evaluate the in-sample hit rates for the no covariates model. All the choice tasks were used in the modeling for the various BYO data upper model runs. For those, no in-sample hit rate could be calculated. That being said, we do not expect notable improvement in the in-sample hit rate as a result of the inclusion of BYO data in the upper model. Neither Sentis & Geller (2010), nor Kurz & Binner (2010), observed improvements in the in-sample hit rates. In the politics data in section 2.2, we observed only a moderate improvement in cell A1 (where data were sparse) and none in cell B.

## 2.4 Dating vs. Politics

So why the difference between the two studies in terms of the impact of BYO data on hit rates? When we compare them directly in Figure 4-1 and 4-2 (and overlook the issue that we don't have a designed holdout task for the dating study), the data seems to be saying that the upper model can be beneficial to prediction for holdout sample, but you can only go as far as the in-sample prediction.

In dating, the no covariate model itself already does a good job of predicting choice. We suspect this is because most women are in agreement about what they want—there is relatively less heterogeneity. So even a uniform prior (i.e., what the average woman wants) does a good enough job of predicting. Having a fuller upper model is a bit like eating organic vegetables: while there is no added nutritional value since we see no improvement in prediction, there is more flavour, as we can better describe and understand the individual woman's preference. This is similar to what was found in earlier papers.

**Figure 4-1**

| Dating | | | | |
|---|---|---|---|---|
| Average Hit Rates on Random Tasks | In-Sample (Canadian) | Holdout Sample (US) | | |
| | No Covariates | No Covariates | Key BYO variables | All BYO Variables |
| Preference only | 53% | 53% | 52% | 52% |
| Preference + "No Buy" | 48% | 43% | 44% | 43% |

**Figure 4-2**

| Congressional Politics | | | |
|---|---|---|---|
| Holdout Task Hit Rate | No Covariates | Demo Covariates | BYO Covariates |
| In Sample (Cell A1) | | | |
| Preference only | 52% | 53% | 56% |
| Preference ⊞ "No buy" | 54% | 55% | 57% |
| Holdout Sample (Cell A2) | | | |
| Preference only | 37% | 48% | 55% |
| Preference ⊞ "No buy" | 32% | 39% | 43% |

The politics study is different. Here there is little consensus on what makes an ideal political platform. A uniform prior (no covariates) does a poor job in predicting preferences in holdout sample. A respondent's political leaning is informative in determining his (and others like his) preference. The BYO information is even more useful. In this case, the upper model is more than just ordinary organic vegetables. It not only tastes better, but also comes with additional nutrients—we have greater heterogeneity, as well as better predictive power (because it is needed here).

## 3. Conclusions

This analysis of three different choice studies confirmed our hypothesis that BYO data from the education section contains valuable information. Using BYO data in the upper model as covariates adds nuance—subtle, yet meaningful variation in the lower model estimates. By capturing individual heterogeneity, we bring out the full flavour of HB.

The answer to our second question—does BYO data help to improve model prediction?—is less clear. Prediction may be improved, but only where it is needed. Our research found improvement where data is sparse, as well as in cases where there is little consensus and greater heterogeneity (lots of disagreement across respondents). The more disagreement, the more opportunity there is for predictive gain by employing BYO questions as covariates.

The use of BYO questions in the upper-level HB model also provides a ready-made solution for generalization to future samples. The BYO questions themselves become the "golden" questions that can be quickly administered to new respondents—allowing researchers to apply the HB model to the new sample without the conjoint exercise.

The BYO data can also be coded as extra choice tasks and be included in the model estimation. As pointed out by one of the audience members during the Q&A session at the conference, rather than comparing performance of models where BYO data are part of the upper model against models with no covariates, or demographic covariates, we could also have compared them to models that included the BYO data as additional choice tasks. This is indeed a good suggestion for future work on this topic.

Jane Tang

Rosanna Mau

Mona Foss

## References:

Eagle, T. (2016) "Upper Level Modeling," Sawtooth Software Turbo-CBC

Kurz, P. & Binner, S. (2010) "Added Value Through Covariates in HB Modeling," Sawtooth Software Conference Proceedings

McCullough, P. R. (2014), "In Search of Covariates that Matter," ART Forum

Pachali, M., Kurz, P., & Otter, T. (2014) "How to Generalize from a Hierarchical Model," ART Forum

Sentis, K. & Geller, V. (2010) "The Impact of Covariates on HB Estimates," Sawtooth Software Conference Proceedings

Tang, J., Grenville, A., Morwitz V., Ülkümen G., & Chakravarti A. (2009), "Influencing Feature Price Tradeoff Decisions in CBC Experiments," Sawtooth Software Conference Proceedings

Tang, J. & Grenville, A. (2010), "How Many Questions Should You Ask in Choice Based Conjoint Studies—Revisited Once Again" Sawtooth Software Conference

Tang, J. & Grenville, A. (2013), "Can Conjoint Be Fun?: Improving Respondent Engagement in CBC Experiments" Sawtooth Software Conference

# SIMULATING FROM HB UPPER LEVEL MODEL

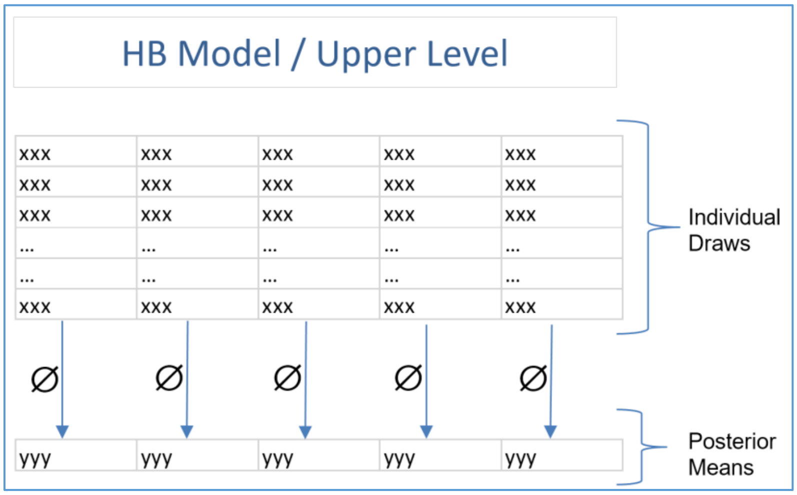
**PETER KURZ**
*TNS INFRATEST*
**STEFAN BINNER**
*BMS MARKETING RESEARCH + STRATEGY*

## Motivation for this Paper

Today many practitioners in market research conduct conjoint analysis or discrete choice modeling (DCM) studies in their day-to-day research work. Since HB became available many years ago to the research community, simulations are quite accurate and therefore many researchers simply base their estimations on standard HB settings and use standard simulation tools. This means that "point estimates," also known as **"posterior means,"** are used in most simulation models. Posterior means are the individual respondents' part-worth utilities, calculated by taking the mean value for each parameter from a certain number of random draws from the posterior distribution after the convergence phase of the HB estimation process. Using posterior means is the standard in Sawtooth Software simulation tools. The disadvantage of this popular simulation method is the risk that due to the averaging of the draws, distribution uncertainty information gets lost. Therefore, simulations based on posterior means might calculate artificial or too simplified preference shares.

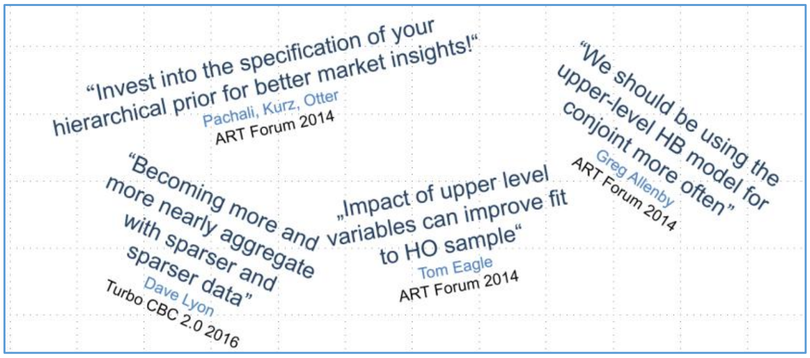
In order to account for both heterogeneity and uncertainty at the individual level, some researchers use individual **random draws** from the lower level model of the HB estimation after convergence is reached and apply those in simulations. Random draws should provide more insights into the uncertainty of respondents' choice behavior and therefore provide more accurate preference shares. One disadvantage of using random draws is that most current standard simulation tools do not support random draw simulations and therefore such simulation tools need to be created individually (e.g., in Excel). Furthermore, the large number of random draws in such tools (e.g., 100 draws for each respondent) leads to very large data sets which might make the simulation tools slow, difficult to operate and sometimes even hard to distribute (e.g., to clients, due to their size).

**Figure 1. HB Model**



Both simulation methods, posterior means and random draws, simulate from the lower level model of the Hierarchical Bayes (HB) model, thus neglecting the upper level model. However, some well-known scientists and research experts emphasize the relevance and impact of the upper level model of HB.

**Figure 2. Expert Quotes**

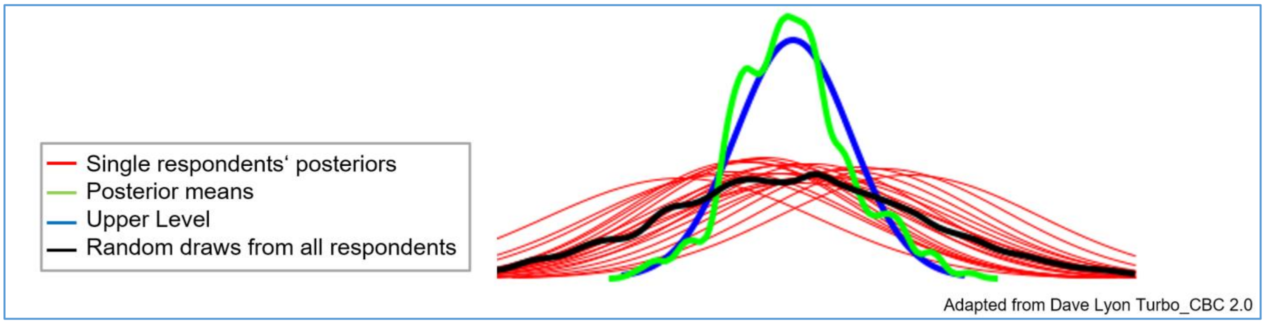

"Invest into the specification of your hierarchical prior for better market insights!"
Pachali, Kurz, Otter
ART Forum 2014
"We should be using the upper-level HB model for conjoint more often"
Greg Allenby
ART Forum 2014
"Becoming more and more nearly aggregate with sparser and sparser data"
Dave Lyon
Turbo CBC 2.0 2016
„Impact of upper level variables can improve fit to HO sample"
Tom Eagle
ART Forum 2014

As there are concerns about losing uncertainty at the individual data, the motivation of this paper is to understand whether simulation results can be improved by using the **upper level model** as the simulation method.

## How HB Works

HB "shrinks" individual-level utilities towards the means of all respondents. This is necessary because often it is impossible to estimate individual respondents—simply because there is not enough information in the data for each single respondent. It is the hierarchical "prior" of HB that pools information across respondents at the population level and allows the calculation of pseudo-individual values and simulations. Therefore, the weaker the individual data, the stronger is the resulting "shrinkage" or smoothing effect of the population level and the results are actually more based on the prior (see Figure 3). The use of posterior means—aggregated mean values of the draws—further strengthens this "shrinkage" effect as it ignores the uncertainty information within individual-level posterior draws (green line in Figure 3 misses the greater variance in the black line).

**Figure 3. HB Shrinkage Effect**



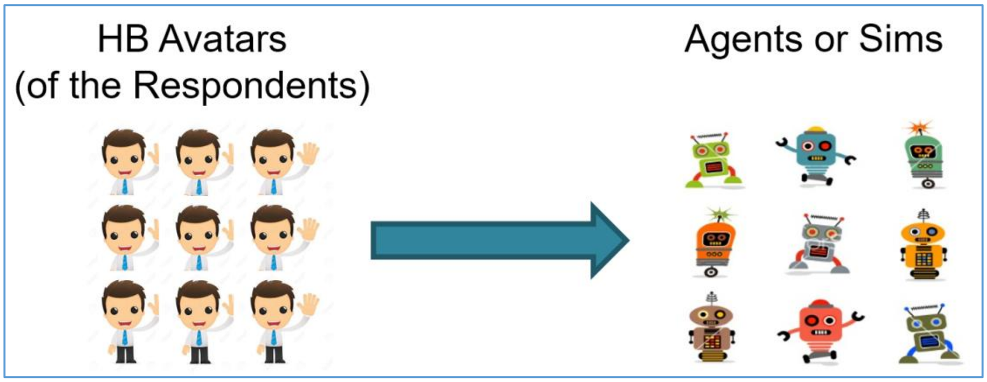
Adapted from Dave Lyon Turbo_CBC 2.0

At the upper level, we assume individuals are distributed in some specified way, usually as multivariate normal, with means and covariances to be estimated. In the lower level, we assume that each individual's answers conform to a separate model, such as logit or regression. Hierarchical Bayes determines the optimal degree to which the upper level model and the lower level model influence the parameters for each individual. The lower level model only dominates if a lot of information per respondent is available.

By applying these hierarchical models, we estimate the population means and covariances at the upper level as well as the part-worths (betas) of each individual at the lower level. The information about the population means and covariances strengthens our estimation of individual results for each respondent. The payoff is that HB permits more precise estimation for each individual, often permitting individual-level estimation where previously only aggregated or segment-level estimation (such as latent class) was possible.

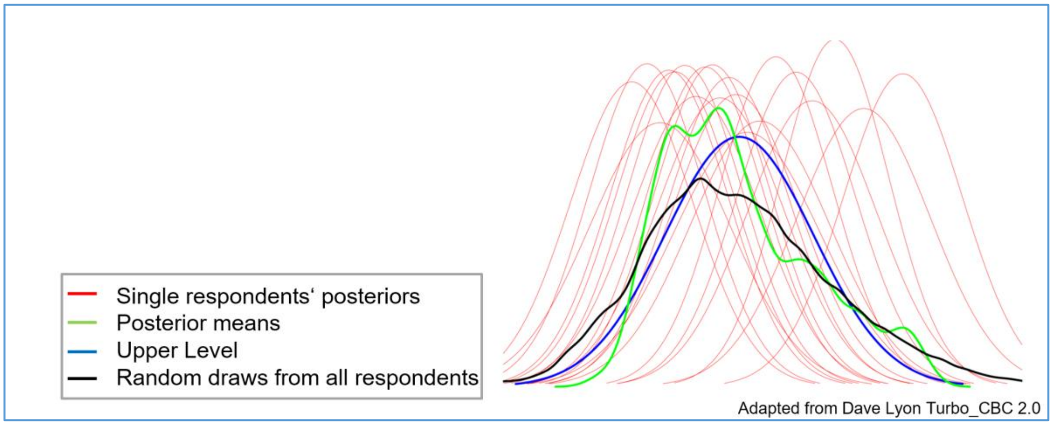
Due to the lack of individual information and the subsequent shrinkage effect, the individual estimates of the lower level model represent “avatars” rather than real respondents. On the other hand, the upper level model allows us to create “agents” or “sims” based on the aggregated functional form which is derived from the respondents in the lower level. The mean values over the avatars and the mean values over the agents are more or less the same.

**Figure 4. Model Characteristics**



The upper level model usually makes the invariant assumption that the data follows a multivariate normal (MVN) distribution (see CBC/HB technical paper, Sawtooth Software 2009). The upper level population means and variance-covariance of the estimates follow that multivariate normal distribution. These parameters are updated in every iteration of the sampler based on draws. The upper level captures the variance as well as the correlation structure in draws at an aggregate level. It is sensitive to the assumption about the functional form (MVN). The covariance matrix characterizes the extent of unobserved heterogeneity. Large diagonal elements, for instance, indicate more (preference) heterogeneity across consumers. Off-diagonal elements indicate patterns in the evaluation of attribute levels (the covariance structure of the part-worth coefficients). For example, positive covariations indicate pairs of attribute levels which tend to be evaluated similarly across respondents. The off-diagonal values can be translated into correlation coefficients. Figure 5 illustrates a model with good representation of the individual heterogeneity by draws and only a small shrinkage effect through the posterior means (the green and blue lines are relatively similar to the black).

**Figure 5. Good Representation of the Individual Heterogeneity**

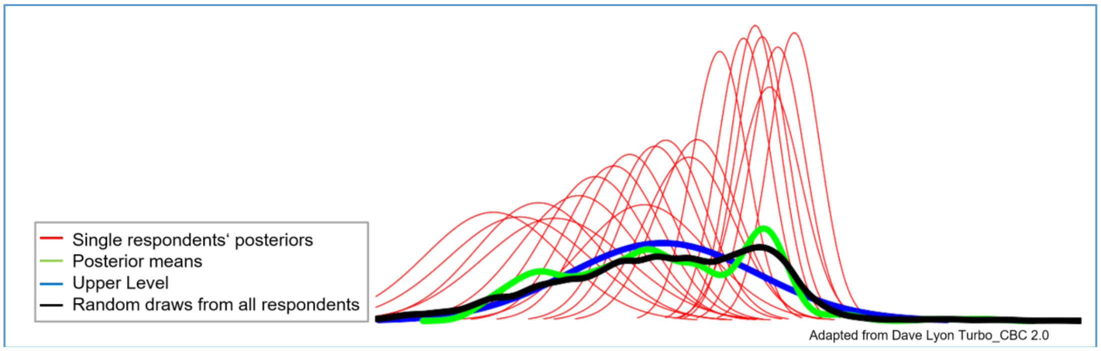


Adapted from Dave Lyon Turbo_CBC 2.0

It is the combination of upper level and lower level models that allows us to estimate the part-worth values. In HB, a part-worth (beta) follows a distribution over respondents. The upper level model contains this full distribution over all respondents. The lower level model identifies the best spot for each individual within that distribution.

Figure 6 shows a different picture: Posterior means, upper level model and draws show similar results which indicates a good overall model fit. However, the plots of single respondents' posteriors show a rather poor representation of individual heterogeneity, especially on the right side of the distribution. This could lead to misinterpretation of the results, if we were—for example—looking at niche segments which are actually not sufficiently covered by the model.

**Figure 6. Shrinkage of Individual Heterogeneity**



Adapted from Dave Lyon Turbo_CBC 2.0

## Upper Level Model—Results

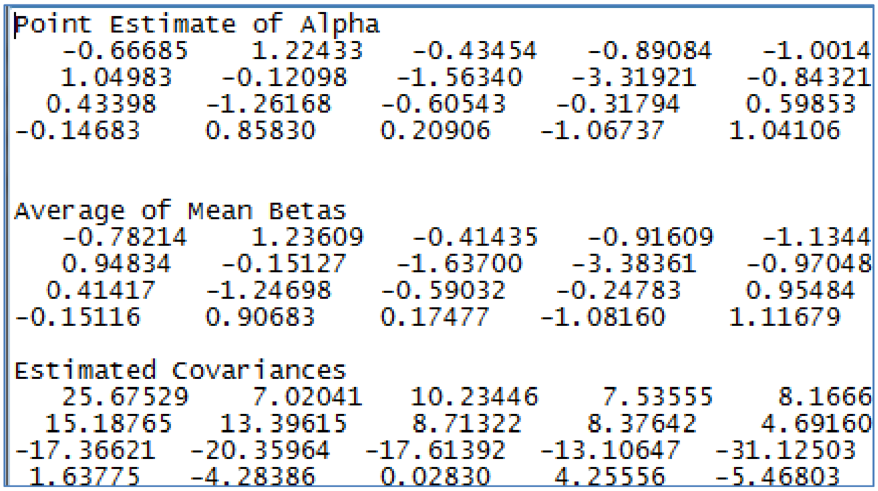
When estimated with the HB sampler, the upper level model has the following aggregated results:

- The "mean value of alphas"—these values are often called the Bayesian logit model, because the alpha values usually come very close to the aggregate logit model. The alphas are the mean values of the population for each attribute level. Mathematically speaking, they are the mean values of the normal distribution of the upper level model.
- The variance and covariance structures—these structures describe the captured heterogeneity and the correlation between the different attribute levels.

Using the above measures offers an adequate representation of the underlying normal distribution of the parameter estimates. (See the blue lines in Figures 3, 5 and 6.)

The described results are part the summary file of CBC/HB:

**Figure 7. Example of a Summary File with the Estimation Results**



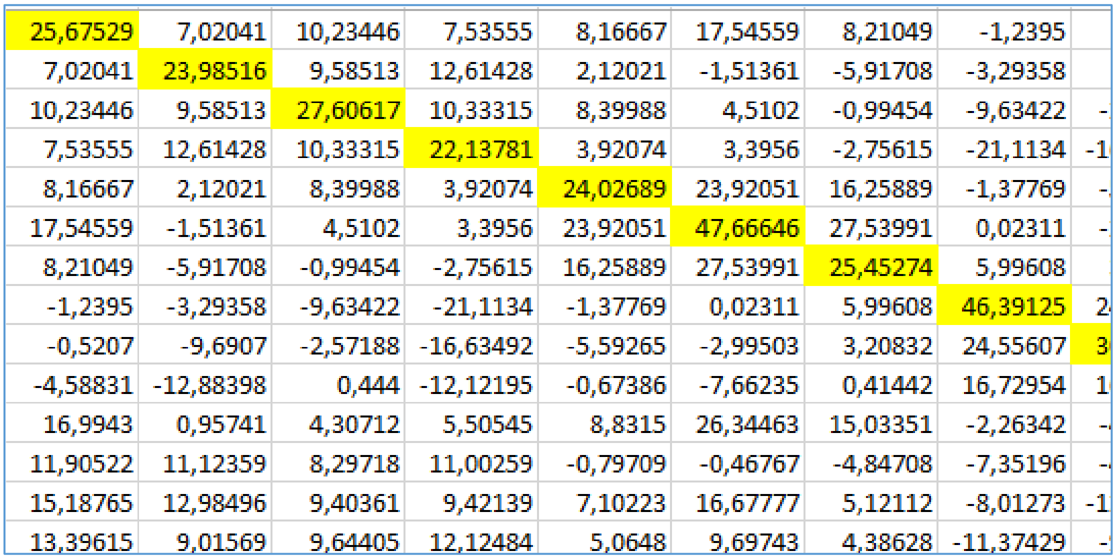
```
Point Estimate of Alpha
   -0.66685    1.22433   -0.43454   -0.89084   -1.0014
   1.04983   -0.12098   -1.56340   -3.31921   -0.84321
  0.43398   -1.26168   -0.60543   -0.31794    0.59853
-0.14683    0.85830    0.20906   -1.06737    1.04106


Average of Mean Betas
   -0.78214    1.23609   -0.41435   -0.91609   -1.1344
   0.94834   -0.15127   -1.63700   -3.38361   -0.97048
  0.41417   -1.24698   -0.59032   -0.24783    0.95484
-0.15116    0.90683    0.17477   -1.08160    1.11679

Estimated Covariances
   25.67529    7.02041   10.23446    7.53555    8.1666
  15.18765   13.39615    8.71322    8.37642    4.69160
-17.36621  -20.35964  -17.61392  -13.10647  -31.12503
  1.63775   -4.28386    0.02830    4.25556   -5.46803
```

By extracting the estimated variance-covariance matrix from the summary file of the estimation results, one could analyze the information contained in the matrix much better.

**Figure 8. Upper Level Variance-Covariance Matrix**



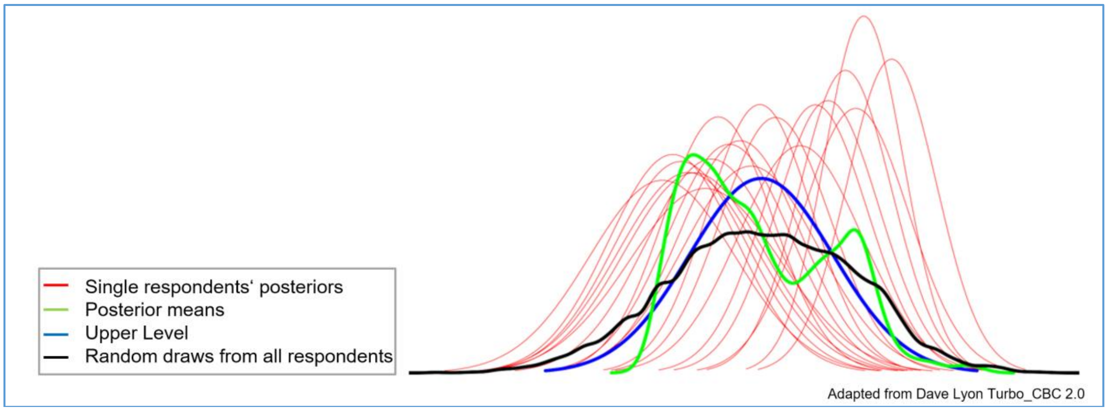
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 25,67529 | 7,02041 | 10,23446 | 7,53555 | 8,16667 | 17,54559 | 8,21049 | -1,2395 | |
| 7,02041 | 23,98516 | 9,58513 | 12,61428 | 2,12021 | -1,51361 | -5,91708 | -3,29358 | |
| 10,23446 | 9,58513 | 27,60617 | 10,33315 | 8,39988 | 4,5102 | -0,99454 | -9,63422 | - |
| 7,53555 | 12,61428 | 10,33315 | 22,13781 | 3,92074 | 3,3956 | -2,75615 | -21,1134 | -1 |
| 8,16667 | 2,12021 | 8,39988 | 3,92074 | 24,02689 | 23,92051 | 16,25889 | -1,37769 | - |
| 17,54559 | -1,51361 | 4,5102 | 3,3956 | 23,92051 | 47,66646 | 27,53991 | 0,02311 | - |
| 8,21049 | -5,91708 | -0,99454 | -2,75615 | 16,25889 | 27,53991 | 25,45274 | 5,99608 | |
| -1,2395 | -3,29358 | -9,63422 | -21,1134 | -1,37769 | 0,02311 | 5,99608 | 46,39125 | 2 |
| -0,5207 | -9,6907 | -2,57188 | -16,63492 | -5,59265 | -2,99503 | 3,20832 | 24,55607 | 3 |
| -4,58831 | -12,88398 | 0,444 | -12,12195 | -0,67386 | -7,66235 | 0,41442 | 16,72954 | 1 |
| 16,9943 | 0,95741 | 4,30712 | 5,50545 | 8,8315 | 26,34463 | 15,03351 | -2,26342 | - |
| 11,90522 | 11,12359 | 8,29718 | 11,00259 | -0,79709 | -0,46767 | -4,84708 | -7,35196 | - |
| 15,18765 | 12,98496 | 9,40361 | 9,42139 | 7,10223 | 16,67777 | 5,12112 | -8,01273 | -1 |
| 13,39615 | 9,01569 | 9,64405 | 12,12484 | 5,0648 | 9,69743 | 4,38628 | -11,37429 | - |

The variances of the attribute level alphas are represented in the main diagonal and explain the heterogeneity of the attribute level, which is the spread around the mean value of the underlying normal distribution. The off-diagonal values reflect the correlation between the attribute levels. For example, the value "7.0241" in above table indicates that Level 3 has a higher preference together with Level 1 compared with for example Level 2. This means that Level 2 is less frequently chosen when Level 1 appears than Level 3.

The hierarchical prior aggregates the information provided by individual-level draws. This "aggregation," however, is sensitive to the specification of the hierarchical prior unless flexible semi-parametric models were considered. This means we capture the overall heterogeneity in general, but not necessarily the functional form of its distribution. As long as we capture most of the heterogeneity, this is only a small loss. However, if the individual draws capture a lot of uncertainty for individual respondents, we lose some of that information about the parameter uncertainty in our model.

Individual draws can be one little step away from a mis-specified model even though they strongly depend on the hierarchical prior if the data is sparse. Among the three, posterior means are least sensitive to the functional form assumed in the hierarchical prior and may produce better aggregate results too, as the following Figure shows:

**Figure 9. Capturing Individual Uncertainty through the Lower Level**



Adapted from Dave Lyon Turbo_CBC 2.0

## Simulating from the Upper Level

Under the assumption that our upper level model parameters are a good representation of the data, simulation from the upper level model would miss a small extent of uncertainty—at the respondent level only. This could be easily compensated for by adding a small extreme value distributed error term. To build a simulator one can simply apply the alphas and the variance/covariance structure, together with this extreme value distributed error term, in order to create a certain number of new “respondents” (or better, “agents”) based on the resulting normal distribution. Based on these new created agents, logit or first choice simulations can be performed with the model. Each agent is used for simulation in the same way as posterior means or random draws would be used.

If the model is based on sparse data and therefore doesn’t capture enough individual respondent behavior (as illustrated in Figure 9), one can add the average within-respondent standard deviation of the draws, instead of the extreme value distributed error term, in order to restore the individual uncertainty—at least approximately.

Remember that each agent comes from the same prior, same scale (variance), same shape (covariance) and same functional form (assumed MVN). However, each agent has a different combination of those—a different random location under the normal distribution.

In contrast to posterior means, both random draws and upper level model agents account for parameter uncertainty caused by sparse data on the individual level. Both depend on the assumption about the functional form of the hierarchical prior. A better understanding of the problem-specific combination of these parameters therefore improves both. The only difference is that random draws are relatively less dependent on the functional form and if data is sparse they show a lot of spread (uncertainty) in the data. Sometimes this information could improve the results, sometimes it could make them worse.

## Empirical Comparison of Simulation Methods

In order to understand the impact of the different simulation techniques we analyzed six real studies (all estimated with standard settings via Sawtooth Software's CBC/HB) and compared the preference shares resulting from these different simulation methods:

1. Posterior Means
2. Random Draws
3. Upper Level Model

For this purpose we randomly selected six real market studies. These empirical studies differ in their complexity (e.g., number of parameters, tasks, concepts per task) as well as sample size, so that one can assume different degrees of sparseness in these datasets.

Furthermore, these studies are different in their research objectives—which could also have an impact on the complexity of the choice tasks. Our sample of studies consisted of:

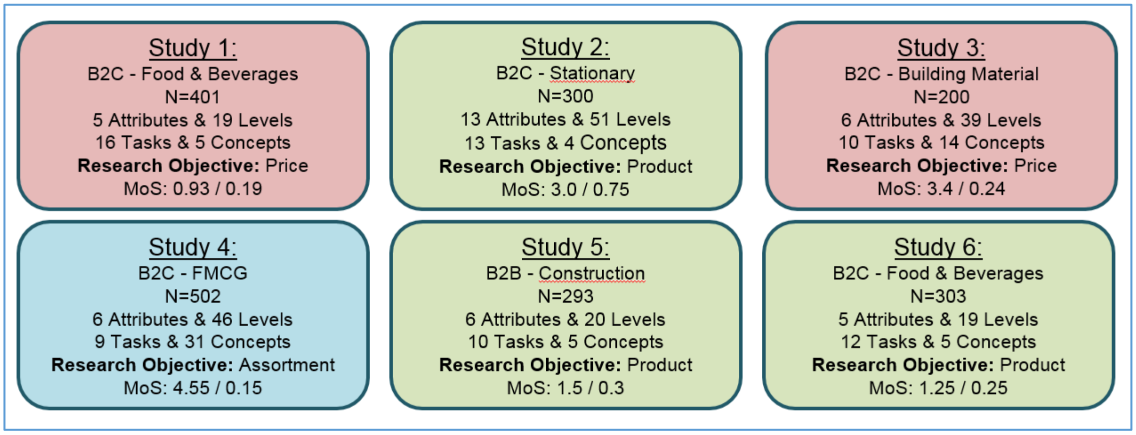
- 3 product configuration studies
- 2 pricing studies
- 1 assortment study

As an indication for expected complexity and sparseness of data we used two measures:

1. Measure of Sparseness 1 (MoS1): $\text{MoS1}=\frac{(\text{\# of Levels} - \text{\# of Attributes}) + 1}{\text{\# of tasks}}$
2. Measure of Sparseness 2 (MoS2): $\text{MoS2}=\frac{MoS1}{\text{\# of concepts}}$

The MoS1 of the six studies ranged between 0.93 (Study 1) and 4.55 (Study 4); MoS2 between 0.15 (Study 4) and 0.75 (Study 2).

**Figure 10. Overview of Six Empirical Studies**



Study 1:
B2C - Food & Beverages
N=401
5 Attributes & 19 Levels
16 Tasks & 5 Concepts
**Research Objective:** Price
MoS: 0.93 / 0.19

Study 2:
B2C - Stationary
N=300
13 Attributes & 51 Levels
13 Tasks & 4 Concepts
**Research Objective:** Product
MoS: 3.0 / 0.75

Study 3:
B2C - Building Material
N=200
6 Attributes & 39 Levels
10 Tasks & 14 Concepts
**Research Objective:** Price
MoS: 3.4 / 0.24

Study 4:
B2C - FMCG
N=502
6 Attributes & 46 Levels
9 Tasks & 31 Concepts
**Research Objective:** Assortment
MoS: 4.55 / 0.15

Study 5:
B2B - Construction
N=293
6 Attributes & 20 Levels
10 Tasks & 5 Concepts
**Research Objective:** Product
MoS: 1.5 / 0.3

Study 6:
B2C - Food & Beverages
N=303
5 Attributes & 19 Levels
12 Tasks & 5 Concepts
**Research Objective:** Product
MoS: 1.25 / 0.25

First we compared the within-sample prediction performance of the three simulation methods. As benchmarks for model performance, we used aggregated hit rates and RLH.

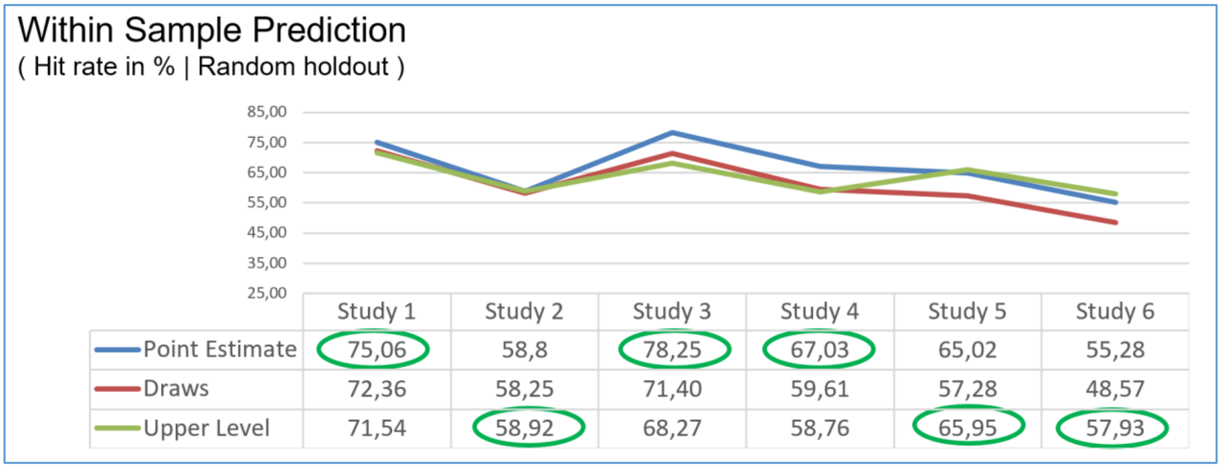
In order to calculate comparable hit rates in each study we used random tasks, which were not included in the estimations.

In regard to random draws, we considered two ways to derive hit rates:

> “0/1 Method”: If more than 50% of the draws of one respondent fit with the random holdout task, this respondent is counted as a hit (if </=50% not).
>
> “100% Method”: Each draw is individually counted as a hit if there is a fit with the random holdout, so one respondent’s hit rate might be 0.6 if, for example, 60 out of 100 draws fit with the random holdout.

As the 100% method provides a higher likelihood to reach hits, hit rates for Random Draws are usually higher when the 100% method is applied. Therefore, we used the 0/1 approach to derive hit rates from random draws simulations.

**Figure 11. Within Sample Prediction—Hit Rates**

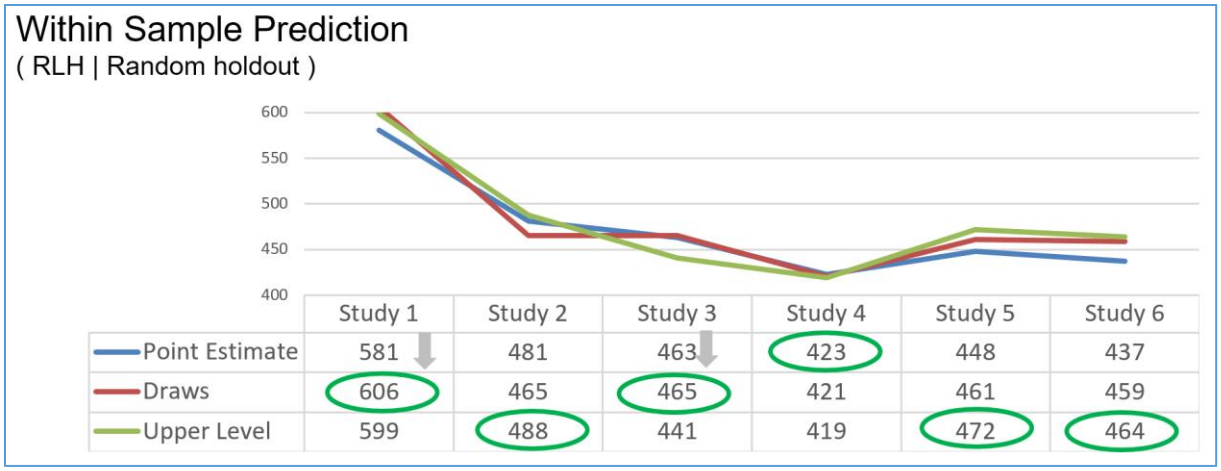


| | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 | Study 6 |
|---|---|---|---|---|---|---|
| Point Estimate | 75,06 | 58,8 | 78,25 | 67,03 | 65,02 | 55,28 |
| Draws | 72,36 | 58,25 | 71,40 | 59,61 | 57,28 | 48,57 |
| Upper Level | 71,54 | 58,92 | 68,27 | 58,76 | 65,95 | 57,93 |

As Figure 11 shows, posterior means and upper level model simulations often lead to very similar results. In the three complex studies focusing on product features (studies 2, 5 and 6), the upper level model simulation performed slightly better than the other two. However, these improvements are quite marginal (and in case of studies 3 and 4, certainly not significant). Nevertheless, a first hypothesis is that the upper level model might improve in-sample predictions compared to posterior means if there is a lot of individual level uncertainty. The rather poor performance of random draws might be caused by the 0/1 holdout method.

The comparison of RLH results shows a quite similar performance between the different simulation methods:

**Figure 12. Within Sample Prediction—RLHs**

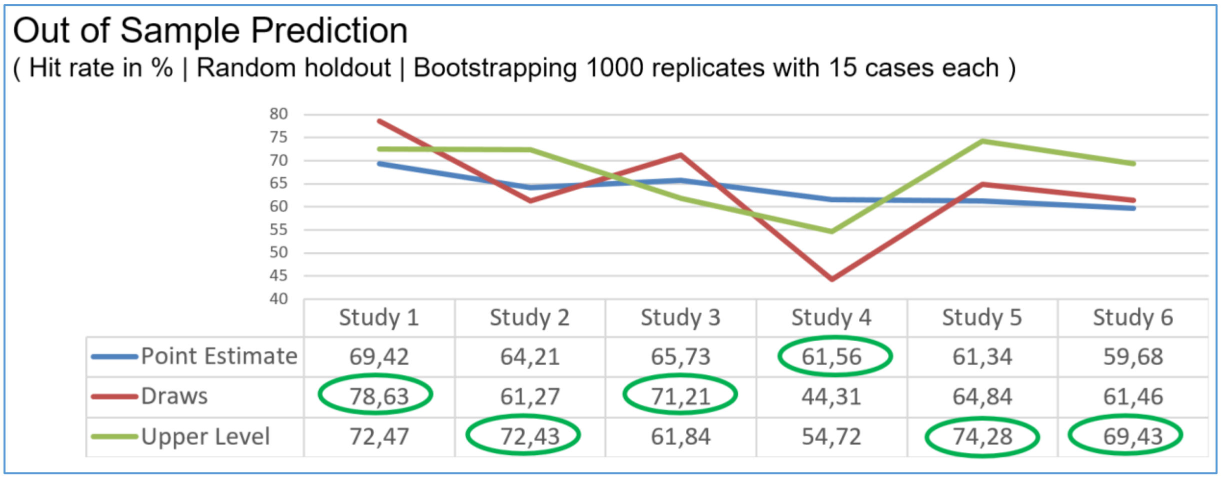


| | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 | Study 6 |
|---|---|---|---|---|---|---|
| Point Estimate | 581 | 481 | 463 | 423 | 448 | 437 |
| Draws | 606 | 465 | 465 | 421 | 461 | 459 |
| Upper Level | 599 | 488 | 441 | 419 | 472 | 464 |

In contrast to the holdout results, the RLH scores showed that the random draws perform slightly better than posterior means in Studies 1 and 3, which were both pricing studies. This is in line with the practical perception of random draws being the simulation model of choice for price-only discrete choice models.

How do the three simulation methods perform in regard to the more relevant measure, the ability of the models to predict *out-of-sample* choice behavior? To provide an indication for out-of-sample validity we used a bootstrapping process: For each of the 6 studies we drew 1,000 different samples, where in each sample 15 respondents were randomly excluded from the estimation and used as holdout respondents. The simulation of these holdout respondents allows us to use both out of sample hit rates and out-of-sample RLH as measures of performance.

**Figure 13: Out of Sample Prediction—Hit Rates**

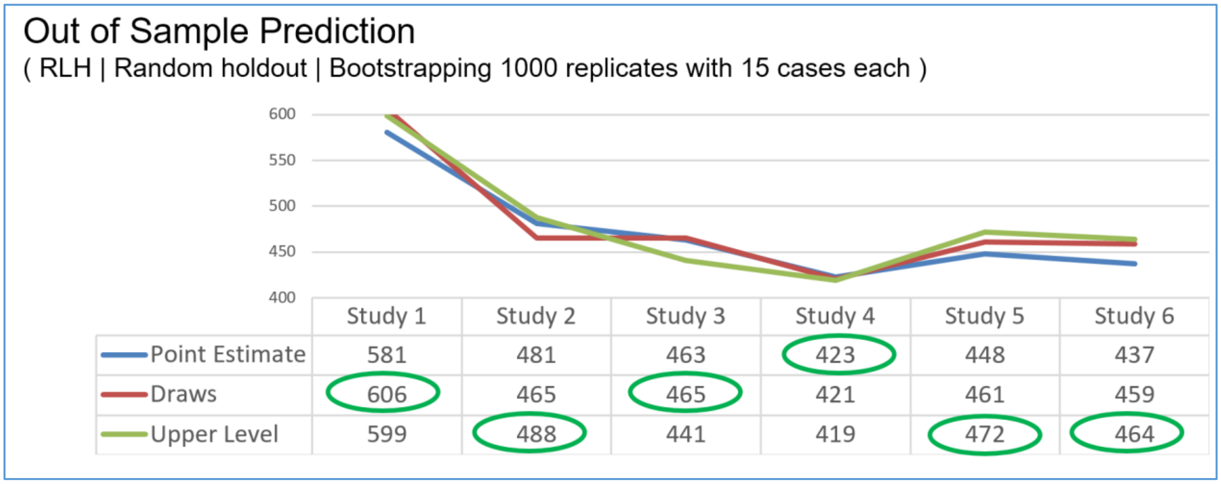


| | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 | Study 6 |
|---|---|---|---|---|---|---|
| Point Estimate | 69,42 | 64,21 | 65,73 | 61,56 | 61,34 | 59,68 |
| Draws | 78,63 | 61,27 | 71,21 | 44,31 | 64,84 | 61,46 |
| Upper Level | 72,47 | 72,43 | 61,84 | 54,72 | 74,28 | 69,43 |

In the out-of-sample holdout prediction the posterior means performed best only in Study 4 (the one with smallest number of choice tasks and an assortment objective). As expected, random draws performed best in the two pricing studies. The upper level model clearly produced the best results for the product configuration studies and outperformed the other two simulation methods.

Similar results can be observed looking at the RLH values:

**Figure 14: Out of Sample Prediction—RLHs**



| | Study 1 | Study 2 | Study 3 | Study 4 | Study 5 | Study 6 |
|---|---|---|---|---|---|---|
| Point Estimate | 581 | 481 | 463 | 423 | 448 | 437 |
| Draws | 606 | 465 | 465 | 421 | 461 | 459 |
| Upper Level | 599 | 488 | 441 | 419 | 472 | 464 |

Again, RLH results are quite close between simulation methods. As in the hit rate results, posterior means performed best in the assortment study, random draws performed best in the two pricing studies and the upper level model clearly produced the best results for the product configuration studies.

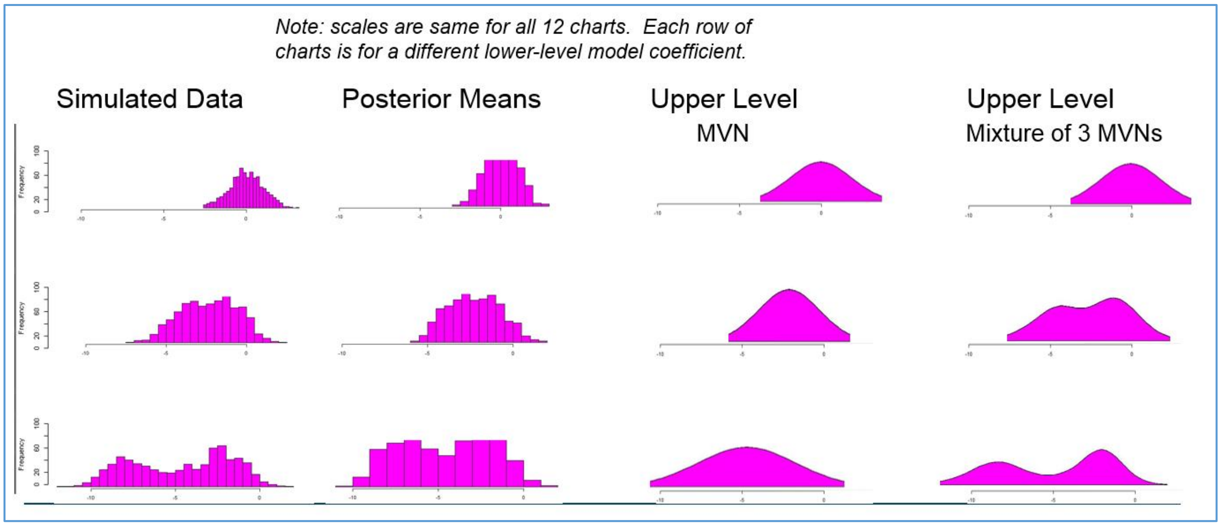
To summarize the performance of different simulation methods in out-of-sample prediction:

1. **Posterior means** only performed best in the assortment study, which had the lowest number of choice tasks and highest number of levels (thus, the highest sparsity of data, with MoS1 = 4.55). Our hypothesis is that there is lot of uncertainty in the lower level which is not real heterogeneity, but rather a reflection of how little we know about the individual respondent. Therefore, the upper level model has no chance to learn much from the respondents and is up to that point mis-specified.
2. **Random draws** performed best in the two pricing studies (even using the 0/1 method). Our hypothesis is that posterior means throw out too much uncertainty here. In the price-only studies we have enough individual information so that the lower level can capture real respondents' heterogeneity and therefore performs better. The simple upper level model we used in this paper was not able to represent the information we captured with the draws.
3. **The upper level model** showed the best results for the studies dealing with product configuration. Here our hypothesis is that there is a lot of uncertainty on the individual level due to the large number of parameters (complex choice tasks). Therefore, posterior means performance is inferior compared to the upper level model. The performance of random draws depends on the way the hit rate is computed (0/1 vs. 100%), but the upper level at least represents the heterogeneity of the respondents in a better way than 0/1 draws.

## Difference between Posterior Means and Upper Level Model

The main difference between posterior means and the upper level model is that the upper level model describes a functional form (the model) of the data while posterior means describe the data by mean values created from a statistical model of the data. The upper level model is therefore more flexible and can be used to sample new agents whenever needed. The only problem is that the upper level model has to be specified more carefully if one wants to use it to build a simulator. Usually the standard HB approach (as in Sawtooth Software's CBC HB) assumes only a single multivariate normal distribution on the upper level model. This could be a significant limitation in estimating the best upper level model, as Figure 15 illustrates:

**Figure 15. Model Comparison**



The three rows show three different lower level model coefficients, the first column the real data (simulated), the second column the posterior means and the third and fourth column the upper level model with two different functional forms. The upper level model is determined by the functional form we have decided to use. As one can clearly see if we use a single MVN for the upper level model, we can capture the normally distributed beta coefficent (first row) very well. The beta coefficient representing a bi-modal distribution with two small peaks (second row) is not recovered very well with the a single MVN upper level model, but a mixture of normals does better. The beta coefficient with a sttrongly bi-modal distribution (row 3) is captured very poorly. Compared to the single MVN upper level model, the posterior means fit the the two bi-modal beta coefficients much better. If we use a more sophisticated upper level model—in this example a mixture of three MVN distributions—the simulation fits very well to all three shapes of beta coefficients.

This underlines that it's really worthwhile to invest more in the proper selection of the *right* upper level model if one attempts to build a simulator based on it. A proper specification of the upper level model is not just a question of the right functional form. Including meaningful covariates could make the upper level model even more powerful. But, we must use really *meaningful* covariates, otherwise the model could be distorted more than improved. Useful covariates are usually exogenous variables, which are related to the attributes and level.

## CONCLUSIONS

Upper level model simulators performed very well—in some cases better than random draws or posterior means, especially when the functional form fits well. In our practical studies this was most often the case in product optimization studies with more attributes when we do not have clear compromise alternative in the choice tasks (Dhar & Simonson 2003). Additionally, in complex studies some attributes or attribute levels may not be taken into account by the respondents. In these cases we assume that the uncertainty captured by the individual respondent draws is due to "attribute non-attendance" (Chrzan & White 2016) and is not useful for predicting the respondents' choice.

Posterior means ignore parameter uncertainty and are shrunk towards the sample mean. This especially leads to poor out-of-sample predictions. In the case of a high "parameters-to-tasks ratio" (sparse data) the upper level model and random draws are superior.

In two-attribute cases (e.g., price-only DCMs), where we usually have clear compromise alternative in our choice tasks, the random draws and upper level model are superior.

The upper level model could be a solution for complex objectives or small sample sizes such as we often find in small markets (e.g., B2B). In such cases the upper level model, which only specifies aggregate parameters and a functional form, often results in more stable and meaningful estimates. An additional advantage of the upper level model approach for such small samples is the ability to resample and rebalance the simulated agents. When it's not possible to have samples that are representative for the market, as is often true in B2B contexts, one can use the upper level model to generate agents that represent the market structure in order to build a better simulator.

In order to cope with high model complexity, investment in the upper level model could be a better solution than longer interviews (Individual Choice Task Threshold, Kurz and Binner 2012). Furthermore the application of meaningful covariates and more complex distributions (such as mixtures of normals) instead of using a standard MVN MNL-HB program could minimize the risk of simulating from a mis-specified hierarchical prior. Working with the upper level model will create a need for HB estimations for choice based conjoint with more customizable computer programs that can vary for each new study.

Although HB methods do not converge to a closed form solution—in the way most of our classical statistical methods do—we should be comfortable with the fact that the variance stabilizes after a few thousand iterations, but there will be still considerable variation in the averages of the parameter estimates. This means that we end up with a distribution of estimates for each individual rather than a single point estimate for each part-worth value.

While the random draws (representing this distribution) are powerful in terms of understanding uncertainty, they add complexity to the analysis. Random draws adequately account for parameter uncertainty but may become impractical for large N data files because we multiply the size of our data files by 100 or even 1000. A well specified upper level model can represent the functional form nearly as accurately as random draws, and is relatively easy to handle because one needs only the aggregate parameters and can easily sample as many agents as necessary on the fly.

## SUGGESTIONS FOR FUTURE RESEARCH:

We often talk about "sparse data" but without a clear understanding of how to define sparseness. Some researchers even talk about sparse data when simulating 2 parameters with 12 choice tasks and 500 respondents, which would mean a MoS1 index of 0.25 (Pachali, Kurz, Otter 2014). It is important to develop a formula or heuristic in order to determine sparse/not-sparse data (similar to the MoS indices) that could be used as a common indicator in the research community. Such a measure could also include parameter/information ratio which would facilitate comparisons among different studies.

Most practitioners use experimental designs which are tested with aggregate logit calculations for d-efficiency only. However, most studies are estimated with HB techniques that take individual respondents into account. Especially if we want to use the upper level model, we should pay more attention to how to optimize the experimental designs for HB techniques.

Should the upper level model be used more often, covariates will become more important for model estimations. Therefore we need good guidance on how to determine meaningful covariates in practice. Up to now, there have been very controversial discussions as to which covariates are meaningful and which ones are not, or could even harm the estimations.

Up to now simulators based on upper level models have been used mainly in the academic world and only a few practitioners have tried to work with them so far. Therefore, we believe that more research has to be done in developing sophisticated but practical upper level model simulators that also take advantage of the resampling and rebalancing capabilities.

Peter Kurz

Stefan Binner

## References

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

Dhar, R., & Simonson, I. (2003): "The effect of forced choice on choice," *Journal of Marketing Research*, 40 (May), 146–160.

Chrzan, K; White, J. (2016): "Mapping Attribute Non-Attendance," *Proceedings of the 2016 Sawtooth Software Conference*.

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

Johnson, R. M. (2000): "Understanding HB: An Intuitive Approach," Sawtooth Software Research Paper Series.

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

Kurz, P.; Binner, S. (2012): "'The Individual Choice Task Threshold' Need for Variable Number of Choice Tasks," *Proceedings of the 2012 Sawtooth Software Conference*.

Kurz, P; Binner, S.(2015): "Capturing Individual Level Behavior in DCM," *Proceedings of the 2015 Sawtooth Software Conference*.

Liakhovitski, D.; Shmulyian, F. (2011): "Covariates in Discrete Choice Models: Are They Worth the Trouble?," 2011 Advanced Research Techniques (ART) Forum presentation.

Lyon, D. W. (2016): "Hierarchical Bayes: Poking Around Under the Hood," presentation at Sawtooth Software Turbo CBC, Captiva Island.

Pachali, M.; Kurz, P.; Otter, T. (2014): "How to Generalize from a Hierarchical Model," 2014 Advanced Research Techniques (ART) Forum presentation

Sawtooth Software (2009): "The CBC/HB System for Hierarchical Bayes Estimation Version 5.0 Technical Paper," Sawtooth Software Technical Paper Series.

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

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

# MAPPING ATTRIBUTE NON-ATTENDANCE

**KEITH CHRZAN**
*SAWTOOTH SOFTWARE*
**JOSEPH WHITE**
*MARITZCX*

Respondents completing Choice-Based Conjoint (CBC) surveys can process attributes in different ways. They may consider all attributes simultaneously, weighing them in their minds, making compensatory tradeoffs and so on, as in the standard random utility theory embodied in the multinomial logit (MNL) model. Other choice theories suggest that respondents consider attributes sequentially: either through multi-step lexicographic models or elimination-by-aspects models or through two-stage models involving a conjunctive or disjunctive whittling of alternatives prior to making a final choice through a compensatory tradeoff decision. Recently attribute non-attendance (ANA) has been suggested as yet another attribute processing strategy respondents may adopt: they may opt to ignore some attributes altogether and to attend only a subset of attributes a researcher shows them (Hensher, Rose and Greene 2005).

ANA has spawned an exploding literature in the fields of transportation economics, health economics and environmental economics, while it has been virtually ignored in the field of marketing research. After reviewing different methods used to identify ANA and after briefly illustrating the practical reason researchers should care about ANA, we re-examine several existing data sets to test how design decisions we make as researchers affect the incidence of ANA among respondents. Finally, we investigate whether using ANA indicators as covariates can improve models built from hierarchical Bayesian MNL analysis.

## WHY ACCOUNT FOR ANA?

By far and away the dominant reason researchers worry about ANA is that it affects willingness-to-pay estimates. Almost all published investigations find that failing to account for ANA can alter conclusions about WTP and most of them find that failing to account for ANA biases WTP upwards (Hensher and Greene 2010, Hensher and Rose 2009, Shen *et al.* 2014). To illustrate this, we took a current project and computed WTP for attributes when we ignored ANA and when we took it into account. To report results for a typical attribute in this is study, storage capacity measured in gigabytes, we found a WTP dollar value of $46/GB when modeled without respect to ANA that reduced to $29/GB when we took ANA into account—a 37% reduction! In our experience WTP calculations often result in exaggerated dollar values and taking ANA into account tends to move WTP estimates in a more realistic direction.

## MEASURING ANA

The earliest papers on ANA measured it with simple respondent self-reports of which attributes, if any, they ignore when making choices (Hensher, Rose and Greene 2005, Hensher 2006). This method, “stated” ANA, raises a variety of questions; for example do we ask the question before respondents complete the CBC exercise, such that that they answer in a vacuum, or do we ask it after they complete the CBC questions? Some researchers have even suggested asking the stated question after each individual CBC question (Scarpa *et al.* 2009a) or asking it about specific attribute levels (Erdem *et al.* 2013). Moreover, the wording of the stated question can make a difference, as asking about which attributes "guide" the decision can yield different results than asking respondents about which attributes they ignore (Scarpa *et al.* 2011).

Regardless of how one measures stated ANA, the next step is to account for it in modeling by removing the effect of attributes a respondent ignores from the utility model for that respondent. When using Sawtooth Software's CBC one would manually recode the design matrix by using a 0 code to denote non-attended attributes. After utility estimation one would also manually recode an individual's non-attended attributes with zero utilities for each level. If using the ACBC software, you could ask the stated ANA question before the ACBC experiment and then prevent non-attended attributes from even entering a respondent's questions by using constructed lists (in which case the 0 coding of dropped attributes for the design matrix and setting the utilities to zero can be handled automatically by the software). Ideally, removing unattended attributes from the analysis in this way eliminates the bias in WTP that ANA can cause.

One can also identify non-attended attributes analytically, based on the choices respondents make. Two methods, one based on latent class MNL and one on mixed logit, account for most models of "inferred" ANA. Campbell *et al.* (2011) propose a latent class model to identify ANA. Assuming an experiment with K attributes which respondents may attend or ignore implies $2^K$ possible combinations of attribute attending and ignoring. The Campbell *et al.* method uses a latent class model with $2^K$ classes where each class constrains a unique subset of the attributes to have utilities of 0.0. For example, a study with variables Q, R and S would have one latent class where all of Q–S are attended, one in which they are all ignored (i.e., set to 0.0), three latent classes with but a single attended attribute and three latent classes with two of the three attributes attended. The highest probability class for each respondent indicates which attributes that respondent ignored and these can be zeroed out as described above before final model estimation. Unfortunately, the latent class estimation becomes very time consuming when there are more than a handful of attributes, or even when any of the attributes have part-worth rather than linear utility functions. The only commercial software package that has automated this approach caps it at a very limiting K=4 parameters.

A mixed logit approach for inferring ANA described by Hess and Hensher (2010) uses the coefficient of variation (CV) from respondent-level logit coefficients and standard deviations (CV is the standard deviation of a parameter divided by the mean parameter estimate). Using CV=2 as a cutoff any coefficients larger than half the size of their standard deviations would count as attended. Selecting CV=2 as the cutoff may seem a little arbitrary but it is easy to describe and implement. This approach will also work for hierarchical Bayesian (HB) MNL. Yardley (2013) uses a different empirical method to identify ANA in HB models, computing cutoffs as percentages (10%–50%) of the largest range for any attribute for a given respondent and finds that larger percentages (at least through 50%) produce better predictions to holdout choice sets. A drawback of either empirical approach to identifying ANA in HB analyses is that it confounds ANA and preference heterogeneity (Hess *et al.* 2013). Using either the Hess and Hensher or the Yardley cutoff methods involves ignoring the distinction between unattended and unimportant attributes. If "non-attended" just means "not very important" then ANA just gives a new name to an old fact (indeed, a white paper posted to the Sawtooth Software website in 2001 notes that respondents with very small price utilities may show impossibly large WTP estimates [Orme 2001]).

Eye-tracking technology provides a third approach for measuring ANA—"visual" ANA. Observing which attributes a respondent's eyes scan should provide a credible, observable measure of ANA (van Loo *et al.* 2014). Eye-tracking suggests that the attributes which respondents attend are the important attributes that make their profiles attractive (Meissner *et al.* 2016). Model fit improves and WTP estimates increase when one accounts for visual ANA (van Loo *et al.* 2014). Curiously, some attributes that do not receive measurable visual attention still end up being important predictors of choice: Olsen *et al.* (2016) report an eye-tracking study where respondents seem able to evaluate levels at which they do not even take a fleeting glance, suggesting that visual ANA may also be problematic.

As one might expect, stated and inferred measures of ANA disagree about which respondents attend which attributes (Campbell and Lorimer 2009, Hess and Hensher 2010, Carlsson *et al.* 2010, Alemu *et al.* 2011). Moreover estimation of ANA from visual ANA differs again from both stated and inferred methods (Balcombe *et al.* 2015, van Loo *et al.* 2014).

## Previous Research on Survey Design Factors Affecting Incidence of ANA

As the different definitions of ANA change our conception of which attributes are ignored by which respondents, ANA appears to be a labile construct, one potentially affected by researcher-controllable decisions about study design. Previous empirical studies bear this out. Hensher (2006) finds that the number of choice sets and of alternatives per choice set affect the incidence of ANA (though Weller *et al.* 2013 find that they do not). Partial profile designs reduce ANA relative to full profile designs (Yardley 2013). Design factors like orthogonality and similarity of attributes can reduce ANA (Cameron and DeShazo 2010) while Bayesian D-efficient designs have higher levels of ANA than do orthogonal designs (Alles and Rose 2014), as do labeled designs compared to unlabeled designs (de Bekker-Grob *et al.* 2010). Finally, designing research to mitigate hypothetical bias through cheap talk scripts or honesty priming can reduce the incidence of ANA (Bello and Abdulai 2016).

Note that we do not suggest that high levels of ANA cause detriment or that low levels provide benefits: ANA may reflect an appropriate simplifying strategy, not an inherently bad thing at all (Nguyen *et al.* 2015). The incidence of ANA does vary quite a bit across studies, however, so in the next section we want to revisit some existing CBC data sets to examine on what other researcher-controllable dimensions ANA may vary.

## Empirical Comparisons—Survey Design Factors Affecting Incidence of ANA

We start our empirical exploration by considering the impact of experimental design decisions on the prevalence of ANA. Specific considerations include Level Overlap, Partial Profile, Best-Worst DCE, Best-Worst Case 2, as well as ACBC relative to standard minimal overlap DCE strategies. Each of these alternative strategies is in one way or another designed to elicit a more thoughtful choice process and consideration of a greater portion of the design space, and as such have potential for reducing ANA.

In what follows we consider a coefficient of variation (CV) of 2 to determine if an attribute was attended or not. Recall from above that CV is simply the standard deviation of a parameter divided by the mean parameter estimate. The CV is calculated using standard output files from Sawtooth Software's CBC/HB application. As noted earlier, the choice of CV is arbitrary, so we chose 2 to be consistent with the literature. The story is invariant with respect to CV choice.

The first strategies we consider are Level Overlap and Partial Profile. ANA and dominant preferences or choices are very much related, and in the extreme if there is a single dominant attribute or level in a design, lesser preferred attributes will be completely ignored, which is to say they are not attended. In the case of level overlap, the inclusion of ties on a dominant attribute should force the decision down to the lesser important factors, thereby eliciting attendance by design. In the case of partial profile this effect is maximized as hidden attributes represent complete overlap and decisions must be made on remaining factors. Through the experimental design therefore, we expect PP to draw attendance to more of the attributes in the design space.

The two studies we have revisited for the level overlap and PP designs are both tablet studies. The design specifications are:

- Tablet Study 1: 6 attributes, balanced versus minimal overlap, presented as triples
- Tablet Study 2: 8 attributes, full profile versus partial profile with 4 attributes presented per task, both FP and PP presented as triples

**Attribute Attendance**

| | Minimal Overlap/FP | Level Overlap/PP |
|---|---|---|
| Tablet Study 1 | 91% | 92% |
| Tablet Study 2 | 72% | 88% |

The results suggest that level overlap may have a small impact on overall attendance, but this may simply be a result of having such a high degree of attendance to start. Because of this the impact of level overlap on ANA is inconclusive in this study. However, in our full versus partial profile strategies we see a very large and significant increase in the percentage of attributes attended.

The next design considerations were Best-Worst DCE and BW-Case 2. Best-Worst DCE is simply a choice experiment where we ask for both the most and least preferred option from a set of alternatives. The idea here is that it is quite possible that what drives preference is different from what folks might avoid. So including the least preferred alternative may entice individuals to consider attributes and levels not playing a role in the identification of the most preferred option.

BW-Case 2 asks questions more like a MaxDiff exercise than a typical CBC task. Respondents are presented with a product profile and then asked to identify which feature/level is most and which is least appealing. Below is an example task.

| Considering the Tablet PC described below, please indicate which one feature you find least appealing and which one feature you find most appealing | | | |
|---|---|---|---|
| | | Least | Most |
| Brand | Samsung Galaxy | | |
| Screen | 10" | | |
| Storage | 32GB | | |
| Memory | 2GB | ✓ | |
| Battery Life | 6 hours | | |
| Price | $199 | | ✓ |

In this example our respondent finds the 2 GB of memory to be least appealing and the price of $199 to be the most. Because we ask people to directly identify the most and least appealing feature of a single profile it is expected that the nature of the task will result in greater levels of attendance.

We looked at two different studies for BW-DCE and BW-Case 2 versus First Choice/CBC, summarized as follows:

- Airline Study: 4 attributes, BW-DCE and BW-Case 2, minimal overlap efficient design
- Refrigerator Study: 4 attributes, CBC and BW-Case 2, also minimal overlap efficient design

The first choice only was used from the Airline Study as our CBC baseline model, from which we then added the worst choice for BW-DCE. Attendance levels are reported in the table below.

**Attribute Attendance**

| | First Choice/CBC | BW-DCE | BW-Case 2 |
|---|---|---|---|
| Airline Study | 82% | 88% | 97% |
| Refrigerator Study | 57% | N/A | 76% |

In our airline study, we see an increase in attendance by including the worst portion of the task, so it appears folks are paying more attention when required to identify a more complete ranking. In this case we presented triples, so best and worst provide a complete ranking. However, the big improvement is seen when we move from the standard CBC to BW-Case 2. In the airline study BW-Case 2 virtually eliminates non-attendance, and dramatically increases attendance in the refrigerator study. It is worth emphasizing that BW-Case 2 is a different kind of exercise, and results in different parameter estimates, so we do not necessarily recommend its use to counter ANA effects, especially if willingness to pay is important.

The final design strategy explored was Adaptive CBC. Adaptive CBC consists of three phases; build your own, screening, and tournament. In the first stage respondents identify their preferred product. The screening phase consists of questions designed to identify levels of attributes that respondents would not consider, thereby creating choice tasks that are more relevant to the individual, i.e., an appropriate consideration set. Finally, respondents go through a series of choice tasks similar to a normal CBC study (except that it is a tournament design where winning concepts return in subsequent choice tasks and losing concepts are dropped, until an overall winner is identified).

A housing study was conducted with both Adaptive CBC and CBC design cells. The design space consists of 10 attributes, with resulting impact on ANA presented in the following table.

**Attribute Attendance**

| | CBC | Adaptive CBC |
|---|---|---|
| Tablet Study 2 | 81% | 90% |

As expected, we see a substantial increase in the percentage of attributes attended. Through ACBC's adaptation to a respondent's preferences, it is likely that a more thoughtful and complete processing of the design space is encouraged.

## Form Factor Impacts on ANA

Two of our studies captured the device type that a respondent used to completed the survey. The relatively small screen size of many smartphones and some tablets may encourage respondents to implement screening rules or otherwise simplify the decision-making process. Simplification rules in turn would likely result in fewer attributes being considered.

The Tablet Study 2 from above along with a hospitality study captured device type. The full profile data for the tablet study were used because of the greater likelihood to observe ANA. The hospitality study was partial profile with 12 attributes displayed 5 at a time as triples. Attendance results are presented below.

**Attribute Attendance**

| | PC | Tablet | Mobile |
|---|---|---|---|
| Tablet Study 2 | 74% | 70% | 76% |
| Hospitality Study | 77% | 76% | 79% |

Mobile responders are no more likely to exhibit ANA in our two studies than other responders, and are at least directionally more attentive. The odd result is the low level of attribute attendance for tablet responders in our tablet study. However, this is likely topic related; tablet owners are less likely to consider alternative brands. These are encouraging results as another piece of evidence that we do not need to worry about the quality of data collected on mobile devices.

## ANA Indicators as Covariates in HB Analysis

As noted, accounting for ANA can have a big impact on willingness to pay estimates. One way to account for ANA is to identify non-attended attributes and zero out the effects in the design matrix. Another possible way to account for ANA is to include a set of indicators as covariates during the HB estimation. One would think that these should be informative covariates, either stated or implied. Again we use CV = 2 for implied indicators, and report the impact on RLH below for implied and stated where available.

**Aggregate Root Likelihood**

| | | ANA as Covariates | | |
|---|---|---|---|---|
| **Study** | **Design** | **None** | **Implied** | **Stated** |
| Tablet | FP-CBC | 0.667 | 0.675 | 0.674 |
| | PP-CBC | 0.567 | 0.575 | 0.571 |
| Hospitality | FP-CBC | 0.571 | 0.585 | N/A |
| Airline | Best only DCE | 0.681 | 0.687 | N/A |
| | BW-DCE | 0.681 | 0.685 | N/A |
| | BW-Case 2 | 0.603 | 0.604 | N/A |
| Housing | CBC | 0.553 | 0.568 | N/A |
| | ACBC | 0.634 | 0.639 | N/A |

While we do see some lift in RLH when we include ANA indicators as covariates, the improvement in internal consistency is truly marginal, and likely not worth the effort. It would likely be better, and more theoretically sound, to leverage the indicators to zero out the corresponding effects coding at the respondent level.

## CONCLUSION

The literature identifies ANA as a potential problem for researchers to address, especially when willingness to pay estimates are among their objectives. Our studies agree with the literature in finding a fair amount of non-attendance, more than a trivial amount for sure: it is clear that some respondents just ignore some attributes. We also find that design decisions can affect attendance rates, particularly the use of partial profile designs and ACBC.

As well we saw that in the studies for which we captured device type there was no impact on attendance based on form factor. This is an encouraging addition to the evidence that mobile responders reliably complete our surveys even when presented with more complex experimental designs.

While the main purpose of this paper is to identify those choices within the control of the researcher that can impact attendance rates, we also considered the inclusion of indicators as covariates. In all cases the inclusion of indicators as covariates does nothing to help the model, at least from an internal consistency perspective as measured by RLH. A practical use we hoped we might be able to make of attendance information turns out not to have been helpful at all.

Finally, it is worth revisiting the potential confound between importance and attendance. While accounting for ANA can have a real impact on WTP estimates, it may well be that non-attendance is only a problem if the attribute not being attended would have an impact on the decision had it been considered. When an attribute is simply unimportant, then we would expect and want the parameters to be consistent with non-attendance. Because of this, it is important to think about the impact of leveraging design strategies aimed at increasing attendance to attributes. Are we inflating the importance of attributes by forcing decisions to be based on uninfluential factors in real world choices? If so, then perhaps non-attendance is not so much a problem that needs to be corrected for in modeling as it is a reflection of perfectly rational decision making processes.

Keith Chrzan

Joseph White

## References

Alemu, M. H, M. R. Mørkbak, S. B. Olsen and C. L. Jensen (2011) "Attending to the reasons for attribute non-attendance in choice experiments," FOI Working Paper, Institute of Food and Resource Economics, University of Copenhagen.

Alles, R. and J. Rose (2014) "Stated choice design comparison in a developing country: recall and attribute non-attendance," *Health Economics Review*, **4**:25.

Balcombe, K., I. Fraser and E. McSorle (2015) "Visual attention and attribute attendance in multi-attribute choice experiments," *Journal of Applied Econometrics*, **30**: 447–67.

Bello, M. and A. Abdulai (2016) "Impact of ex-ante hypothetical bias mitigation methods on attribute non-attendance in choice experiments," *American Journal of Agricultural Economics Advance Access*, doi: 10.1093/ajae/aav098.

Campbell, D., D. A. Hensher and R. Scarpa (2011) "Non-attendance to attribute in environmental choice analysis: a latent class specification," *Journal of Environmental Planning and Management*, **54**, 1061–1076.

Campbell, D. and V. S. Lorimer (2009) "Accommodating attribute processing strategies in stated choice analysis: do respondents do what they say they do?" paper presented at the European Association of Environmental and Resource Economists Annual Conference, Amsterdam.

Carlsson, F., M. Kataria and E. Lampi (2010) "Dealing with ignored attributes in choice experiments on valuation of Sweden's environmental quality objectives," *Environmental and Resource Economics*, **47**, 65–89.

Cameron, T. A and J. R. DeShazo (2010) "Differential attention to attributes in utility-theoretic choice models," *Journal of Choice Modeling*, **3**, 73–115.

de Bekker-Grob, E., L. Hol, B. Donkers, L. van Dam, J. D. F Habberma, M. E. van Leerdam, E. J. Kuipers, M-L. Essink-Bot, E. W. Steyerberg (2010) "Labeled versus unlabeled discrete choice experiments in health economics: an application to colorectal cancer screening," *Value in Health*, **13**: 315–323.

Erdem, S.,D. Campbell and A. R. Hole (2013) "Attribute-level non-attendance in a choice experiment investigating preferences for health service innovations" paper presented at the International Choice Modeling Conference, Sydney.

Hensher, D. A. (2006) “How do respondents process stated choice experiments? Attribute consideration under varying information load,” *Journal of Applied Economics*, **21**: 861–878.

Hensher, D. A. and W. H. Greene (2010) “Non-attendance and dual processing of common-metric attributes in choice analysis: a latent class specification,” *Empirical Economics*, **39**, 413–426.

Hensher, D. A. and J. M. Rose (2009) “Simplifying choice through attribute preservation or non-attendance: implications for willingness to pay,” *Transportation Research Part E*, **45**, 583–590.

Hensher, D. A., J. M Rose and W. H. Greene (2005) “The implications on willingness to pay of respondents ignoring specific attributes,” *Transportation*, **32**, 203–220.

Hess, S., A. Atathopoulos, D. Campbell, V. O’Neill and S. Causssade (2013) “It’s not that I don’t care, I just don’t care very much: confounding between attribute non-attendance and taste heterogeneity,” *Transportation*, **40**, 583–607.

Hess, S. A. and D. A. Hensher (2010) “Using conditioning on observed choices to retrieve individual specific attribute processing strategies,” *Transportation Research Part B: Methodological*, **44**, 781–790.

Meissner, M., A. Musalem and J. Huber (2016) “Eye-tracking reveals processes that enable conjoint choices to become increasingly efficient with practice,” *Journal of Marketing Research*, **53**: 1–17.

Nguyen, T. C., J. Robinson, J. A. Whitty, S. Kaneko and N. T. Chinh (2015) “Attribute non-attendance in discrete choice experiments: a case study in a developing country,” *Economic Analysis and Policy*, **47**, 22–33.

Olsen, N. B., K. Uggeldahl, C. Jacobsen and T. H. Lundhede (2015) “Measuring attribute non-attendance in stated choice experiments using statements, inference and eye-tracking.” Paper presented at the International Choice Modeling Conference, Austin.

Orme, Bryan (2001) “Assessing the Monetary Value of Attribute Levels with Conjoint Analysis: Warnings and Suggestions,” Sawtooth Software Research Paper downloaded from https://www.sawtoothsoftware.com/download/techpap/monetary.pdf on October 20, 2016.

Scarpa, R., T. J. Gilbride, D. Campbell and D. A. Hensher (2009a) “Modelling attribute non-attendance in choice experiments for rural landscape valuation,” *European Review of Agricultural Economics*, **36**, 151–174.

Scarpa, R., S. Notaro, R. Raffaelli and J. Louviere (2011) “Modelling attribute non-attendance in best-worst rank ordered choice data to estimate tourism benefits from Alpine pasture heritage,” paper presented at the EAAE 2011 Congress, Zurich.

Shen, M., Z. Gao and T. Schroeder (2014) “Attribute non-attendance in food choice experiments under varying information load,” paper presented at Agricultural and Applied Economics Association’s 2014 AAEA Annual Meeting, Minneapolis.

van Loo, E. J., Nayga Jr, R. M., Seo, H. S., & Verbeke, W. (2014). Visual Attribute Non-Attendance in a Food Choice Experiment: Results From an Eye-tracking Study. Selected Paper prepared for presentation at the 2014 AAEA Annual Meeting, Minneapolis.

Weller, P., M. Oehlmann and J. Meyerhoff (2013) "On the influence of dimensionality of stated choice experiments on attribute-non-attendance," paper presented at the International Choice Modeling Conference, Sydney.

Yardley, D. (2013) "Attribute Non-Attendance in Discrete Choice Experiments," *2013 Sawtooth Software Conference Proceedings*, 195–204.

# USING DISCRETE CHOICE TO HELP INDIVIDUALIZE CUSTOMER LIFETIME VALUE

*MICHAEL SMITH*
*MICHAEL REMINGTON*
*MICHAEL DRAGO*
*THE MODELLERS*

## ABSTRACT

Given the recent rise in demand for creating customer segments based on customer lifetime value (CLV), we helped a professional sports team to segment and type their database into CLV tiers by incorporating actual customer data alongside a discrete choice model to project future spending behavior.

## DEFINITION OF THE PROBLEM AND THE PROPOSED SOLUTION

The ultimate goal of calculating Lifetime Value is to quantify the value individuals represent to an organization over a given period of time. The idea has been largely promoted by Peter Fader. Traditionally it is used in a contractual relationship, but more and more clients are interested in the future value of other customer groups. Market segmentations often focus on attitudes and behaviors, but one of the key questions that is often left unanswered by a segmentation is the actual value a segment represents in the future.

The traditional formula for CLV is as follows:

$$\boldsymbol{Customer\ Lifetime\ Value}\ (\$) = \boldsymbol{Margin}\ (\$) * \frac{\boldsymbol{Retention\ Rate}\ (\%)}{[\mathbf{1} + \boldsymbol{Discount\ Rate}\ (\%)] - \boldsymbol{Retention\ Rate}\ (\%)}$$

The CLV formula has only three parameters: (1) constant margin (contribution after deducting variable costs including retention spending) per period, (2) constant retention probability per period, and (3) constant discount rate per period. Furthermore, the model assumes that in the event that the customer is not retained, they are lost for good. Finally, the model assumes that the first margin will be received (with probability equal to the retention rate) at the end of the first period.

Customer Lifetime Value has an intuitive appeal in marketing because it provides a baseline for each customer that determines how much a marketing department should be willing to spend to win that individual's business. The CLV calculation in a contract scenario is benefited from having a consistent margin of revenue over time. We can also naturally assume that the retention rate and discount rates are constant year over year. One of the questions we have to consider once that constant margin assumption, specifically, disappears is how to develop a framework in which we can estimate the margin for each of the years in question. The solution to this question that we settled on was to use a Discrete Choice Model to calculate yearly margins. We are then able to rewrite our formula to account for yearly changes and potential variability in behavior.

## CASE STUDY

Recently a professional sports team requested that a lifetime value calculation be incorporated into a segmentation study. The goal was to provide their marketing and sales team a more powerful resource in all their marketing activities.

Since sports fans have no contractual agreement with their favorite sports teams their yearly spend can vary greatly year over year, so to calculate the lifetime value we rewrote the CLV formula and developed a Discrete Choice task to explore fan spending. The actual financial benefit of a given respondent will be realized through multiple different channels and their actual expenditures with the team could change dramatically based on a variety of factors. The power of the discrete choice model is that we can explore all of these possible scenarios and see how each individual will respond to the different factors that could come into play.

The factors that were considered in the discrete choice task itself are attributes relating to season expectations and game day experience as well as everything related to tickets. Questions were also asked outside of the choice task to establish demographics, preferences and behaviors as they relate to the team. Combinations of all these factors are used to create a plausible future season. The full list of attributes is contained in Table 1.

**Table 1. Case Study Discrete Choice Attributes**

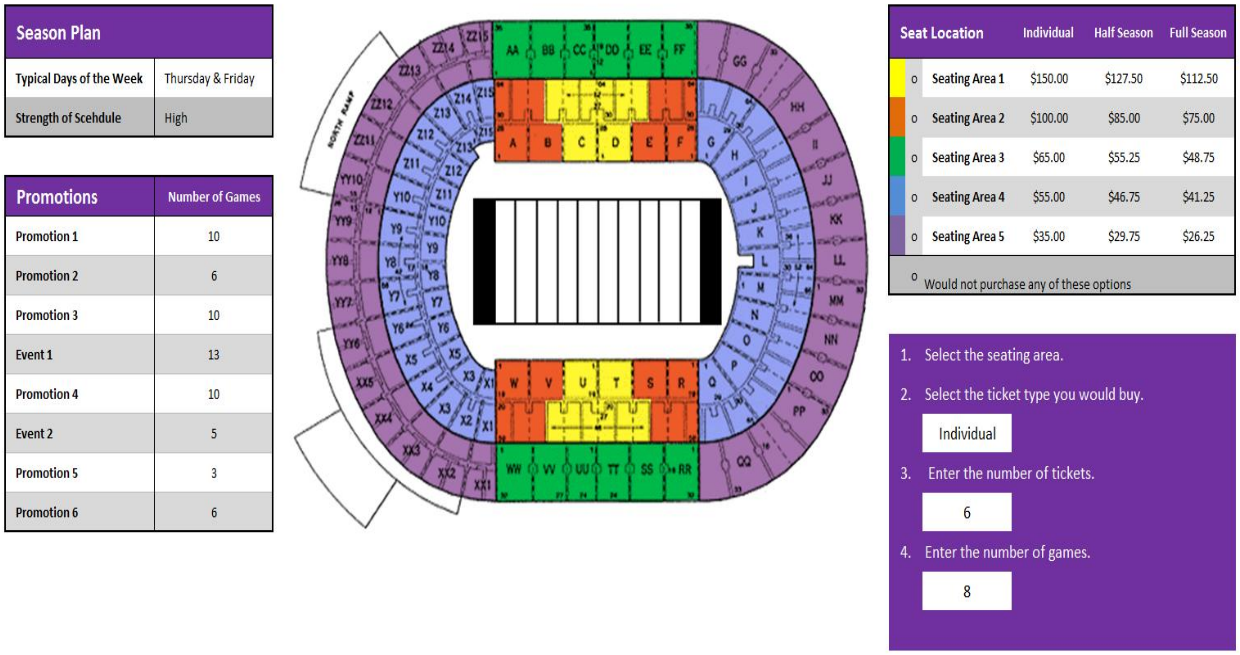
| Attribute Name | Level1 | Level2 | Level3 | Level4 | Level5 | Level6 |
|---|---|---|---|---|---|---|
| Price per Ticket | 70 | 55 | 40 | 30 | 20 | |
| Package Type | Single Game | Half Season | Full Season | | | |
| Typical Days of Week | Wednesday, Friday, Saturday | Friday & Saturday | Thursday & Friday | Thursday, Friday, Saturday | Saturday & Sunday | Friday, Saturday, Sunday |
| Strength of Schedule | Low | Medium | High | | | |
| Promotion 1 | 13 | 11 | 7 | 4 | 2 | |
| Promotion 2 | 12 | 8 | 6 | 4 | 2 | |
| Promotion 3 | 15 | 11 | 9 | 7 | 0 | |

Using these attributes we are able to present various potential season configurations to the respondents and gauge their response. Figure 1 below shows an example of the task.

The task revolves around a stadium map. The key question put to a respondent is given the various levels of the attributes for this potential season, they must choose where in the stadium they will sit, the type of tickets they would buy and how many. We are able to present various ticket prices that are all scaled up or down based on the seating area, as well as a variety of promotional and event information as well. Knowing the number of tickets that any given respondent will purchase is key to calculating the lifetime value, especially because there are different tiers of ticket pricing as well as different ticket packages to choose from.

Once the data is gathered we used a volumetric model to predict percent of maximum spend. Essentially we are predicting how much of the maximum they would be willing to spend given a season with a certain combination of the attributes and levels. Their favorite season offering where they would spend the most money would be coded to a response of 1 or 100% of their maximum and all other tasks would be coded to be a percent of this maximum spend number. The goal in the volumetric model is to not only gauge which season configuration is most preferred, but also which configurations would in fact result in the greatest revenue. In our case our task was very focused on tickets, but in every case it is important to consider what the potential revenue sources are to adequately capture all the sources of revenue. Missing individual level revenue sources will result in less accurate predictions that will trickle down to every application of the Lifetime Values after they are calculated.

**Figure 1. Example Choice Task**



| Season Plan | |
|---|---|
| Typical Days of the Week | Thursday & Friday |
| Strength of Scehdule | High |

| Promotions | Number of Games |
|---|---|
| Promotion 1 | 10 |
| Promotion 2 | 6 |
| Promotion 3 | 10 |
| Event 1 | 13 |
| Promotion 4 | 10 |
| Event 2 | 5 |
| Promotion 5 | 3 |
| Promotion 6 | 6 |

| | Seat Location | Individual | Half Season | Full Season |
|---|---|---|---|---|
| 🟨 | ○ Seating Area 1 | $150.00 | $127.50 | $112.50 |
| 🟧 | ○ Seating Area 2 | $100.00 | $85.00 | $75.00 |
| 🟩 | ○ Seating Area 3 | $65.00 | $55.25 | $48.75 |
| 🟦 | ○ Seating Area 4 | $55.00 | $46.75 | $41.25 |
| 🟪 | ○ Seating Area 5 | $35.00 | $29.75 | $26.25 |
| | ○ Would not purchase any of these options | | | |

1. Select the seating area.
2. Select the ticket type you would buy.
   Individual
3. Enter the number of tickets.
   6
4. Enter the number of games.
   8

The simulation output of the model can be translated back to actual dollar figures. The maximum they indicated they would be willing to spend is multiplied by the percent of maximum predicted by the model for each of the 10 years that we look at. This of course factors in all the possible sources of revenue that we studied. These amounts of spend are ready to be used in the restructured Customer Lifetime Value calculation. There are a few changes that need to be made from the original formula for it to work in this framework.

In order to incorporate the Discrete Choice data into the Customer Lifetime Value calculation it needs to be rewritten in a general format. The new general format no longer assumes margin, retention rate or discount rate are constant per time period. Instead it is written as follows:

$$\boldsymbol{Customer\ Lifetime\ Value}\ (\$)\ = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{Margin_i}\ (\$)\ * \boldsymbol{Retention\ Rate_i} * (\boldsymbol{1} - \boldsymbol{Discount\ Rate_i})$$

“N” indicates the number of years, or number of time periods in question. Margin becomes the revenue projected from the DCM simulator for each year represented by “i.” This figure is a dollar figure predicted by the model for year$_i$ based on how the season will be configured. In the rewritten formula it's important to note that unlike the simplified formula where retention rate and discount rate are a compounded figure to calculate value in years further out, the rates in this formula are slightly different. They are per year estimates of rates that are used to calculate a net figure for a given time frame. In order to determine reasonable retention rates we gathered a variety of information that we then used to calculate per year estimates of how likely a respondent was to remain a customer of the team. The retention rate is variable, because each successive year into the future a fan becomes less and less likely to stay. Since the factors that influence a fans likelihood to stay are many, just like with sources of revenue, it is important to consider what factors might influence that likelihood and gather the appropriate data to make the appropriate projections. One final piece of this equation is that since data projections get less certain the further into the future they are projected it is important to work with the client to determine a reasonable discount rate for each year going forward. Once we know the margin, retention rate and discount rate for each year, we can then solve the equation and sum up the years to get the lifetime value for a respondent.

This equation is the key to the whole process. This is what gives us the power to estimate lifetime values for any specified timeframe. We may not have the non-varying numbers that provide a simpler framework like we do when we work with contracts, but with our simulator in hand we can project out a lifetime value for our given time frame, in this case 10 years. Having this input on an individual level means that for any possible combination of the levels tested we can predict on an individual level what value that customer represents to the team. This opens the door for a variety of next steps we can take to provide greater value in understanding the customer base.

The lifetime value of each respondent can be used as an additional input to the segmentation. The benefits of the segmentation are that not only will it be clear what value a certain group represents to the team, but the team will also have in-depth profiles informing them on a wide variety of variables to help them understand who the group contains. This segment can also be used in a simulator to look at the discrete choice data to study what motivated this particular segment's decisions and how to maximize revenue from the group. The team could even make decisions about promotions and events based on what will prompt the largest gains from their target segments.

Another benefit that is gained from adding the Discrete Choice data into the Customer Lifetime Value calculation is that the team also has the ability to plan future seasons. Since the appeal of each of the attributes and levels is known, it is easy to determine the best season configurations for different potential customer segments. It's even possible to return and design specific season configurations geared around the segments that were created based on the Customer Lifetime Value. Doing that, the team can even further maximize their ability to target each group.

One final step that we took was to create a typing tool that can be used in the team database to classify everyone they have gathered information about. The team boosts their ability to find and target key customer groups even further. The team can continue to leverage the vast amount of information provided to them by the Discrete Choice Model and subsequent segmentation as they look at new potential customers and understand what will motivate each group to bring in more revenue to the team.

These are only three analyses we provided to our client, but there are likely many more possible applications for the Customer Lifetime Value calculation when it is based on data from a Discrete Choice Model.

Discrete Choice models provide a wealth of power in understanding future spending habits of individuals as opposed to an aggregate view. The team's future plan provides clarity into expected retention in the future. Individuals can be divided into clearly defined groups by incorporating estimated value represented to the organization over the given time. The profiles of these segments provide the organization with an actionable database that will guide marketing activities to be more customer specific and focused on the right customers.

Michael Smith

Michael Remington

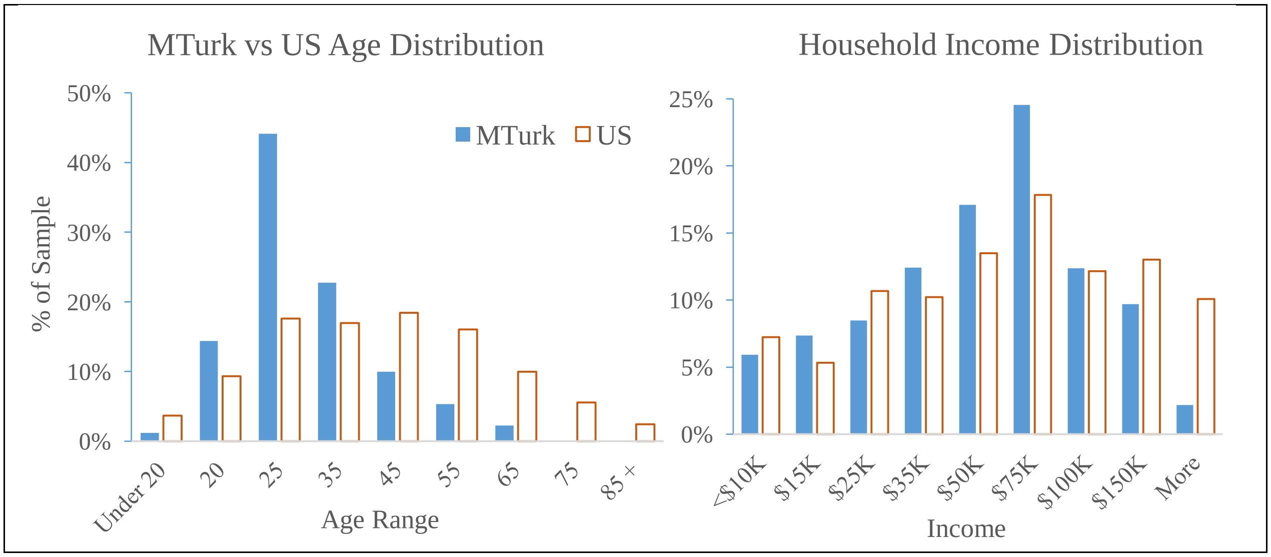
Michael Drago

# MTurk Survey Deception: Sources, Risks, and Remedies

**Kathryn Sharpe Wessling**
*Wharton School*
**Joel Huber**
*Duke University*
**Oded Netzer**
*Columbia University*

Academics and businesses are increasingly using crowdsourcing platforms (e.g., CloudFlower) to obtain inexpensive survey responses. Amazon Mechanical Turk ("MTurk"), with more than a half million registered "Turkers," dominates this space. Founded in 2005 and named after the "The Turk," an 18th century "machine" that was apparently doing human tasks (i.e., playing chess) without human involvement (which was eventually discovered to be a hoax given there actually was a person secretly moving the pieces), Amazon's launch was an attempt to farm out machine-like tasks (e.g., image categorizing) to people who could actually do it better than computers. While not originally intended to be a survey participant pool, its large base of potential respondents (mainly US residents due to tax laws) makes it extremely attractive to researchers. Despite the massive number of registered users, it is estimated that approximately 7,300 survey takers are active at any given time (Stewart *et al.* 2015). Not surprisingly, these online participants tend to be younger than the general U.S. population and have lower household incomes. However, researchers often use MTurk over other web-based participant pools because of the price and speed advantages it offers. A 10-minute survey, paying $1.00 per respondent, typically takes less than four hours to obtain 150 responses.

**Figure 1. Turkers Tend to Be Younger and Poorer than the Typical US Population**



Sources: U.S. Census Bureau, 2014 Population Estimates MTurk: Authors' MTurk Panel Surveys

The increase in crowdsourcing participant pools has also led to an increase in professional survey takers. Numerous papers have demonstrated that the quality of data from MTurk is comparable to more expensive panel services (Weinberg, Freese, & McElhattan 2014). However, other research has found reduced effect sizes amongst nonnative respondents on MTurk (Chandler *et al.* 2014, 2015), suggesting that research conclusions may not significantly differ, but the level of significance will be reduced, hence requiring a greater sample size to obtain the same level of significance.

Much of the research that has examined the quality of MTurk data is based on a general sample without the need to screen or target a specific segment of the population. We will show that given the opportunity to do so, anonymous respondents in crowdsourcing marketplaces deceive on screening requirements to monetarily benefit, and that this deception can distort market research results. We first identified this problem when we ran two health related surveys—one which screened for athletic people under the age of 35 and the other which screened for cigarette smokers over the age of 50. We found that 17% of our respondents passed the screeners in both surveys, an impossible outcome if these respondents were honest and paying attention. Since it is not uncommon that a market researcher may want to qualify respondents based on a specific demographic or behavior, we were interested in systematically measuring the extent of impersonation on MTurk. Most importantly, we wanted to determine if the data obtained from those who deceive differed statistically from those who honestly qualified.

To meet these objectives, we created a panel of Turkers who would serve as the basis for our research. These 1,109 panelists filled out an extensive survey where they indicated basic demographics, personality measures, and possible correlates to deception such as religious practices, materialism, political bent, and responses to moral choices (Graham & Haidt 2012). We also asked a series of questions about current ownership of sports equipment, pets, and technical devices, and their participation in specific online MTurk chatrooms. Because in the panel creation survey there was no financial benefit for respondents to answer in any particular manner, the Turkers' responses to the demographic and product ownership questions in the panel creation survey served as a benchmark to evaluate impersonation in follow-up surveys.

## Deception Experiments

We ran four experiments which were only open to our pre-surveyed panelists. These experiments tested both the degree of impersonation and whether those who impersonate provide different responses to subsequent questions compared with legitimately qualified respondents.

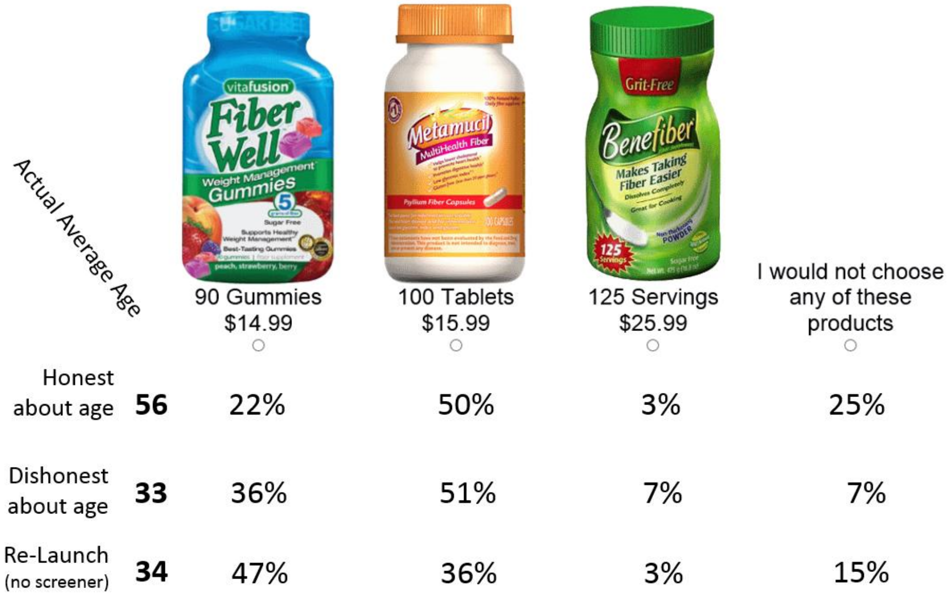
Our first survey paid 75 cents for those who owned a kayak (as indicated in the screener question) and completed the survey. In the initial panel survey, 7.6% of the 1,109 panelists stated that they owned a kayak. In this screener survey, two months later, 88% of the paid participants (who passed the screener) contradicted their earlier statement (that they did not own a kayak). While 6% stated that they had purchased a kayak in the last six months, this left 82% of respondents who indicated a clear discrepancy between their panel survey response (claimed to not own a kayak) and this screener survey response (claimed to own a kayak). For comparison purposes, we relaunched the survey without screening out any respondents (but excluding those who had previously taken the screener version of the survey) and found that only 4% were inconsistent between their panel survey responses and future responses when there is no incentive to lie. Thus, we do not believe that the 82% inconsistency (in the screener survey) is just due to careless responses or measurement error.

Our second survey screened participants on pet ownership before a relatively lengthy pet food survey for $1.25. To pass the screener, respondents needed to have both a dog AND a cat. Of our panelists, 19% had reported in the original panel survey that they had both. Comparing the response on the screener to the information provided in the panel survey, 71% of Turkers who stated that they had both types of pets had previously disclosed in the panel survey otherwise. Of these probable deceivers, 32% had reported not having either a dog or a cat; whereby, the remaining 68% had either a dog or cat, but not both.

These two studies demonstrate that participants will falsify answers to screener questions in order to financially gain. It is unclear, however, if deceivers statistically differ from those who honestly qualify in their responses to the questions in the main body of the survey. We investigate this issue in the next two surveys.

Our next survey screened on age, paying 60 cents to anyone 50 years old or older to be a part of a survey about dietary fiber supplements. Additionally, participants were asked to make a product choice based on a set of fiber brands (Metamucil, Benefiber, or Fiber Well at $14.99, $15.99, and $25.99, respectively). The product choice task also included a “none” option for those who would not choose any of these product offerings. In this survey, 40% of our paid respondents, claimed to be 50 years old or older but had reported being under 50 years old in the original panel survey (as indicated by stated age as well as year and month of birth). These impersonators (average age of 33 years old) significantly differed on the fiber choice task compared to the target group of respondents (those who consistently reported being over 50 years of age). In particular, the shares of Fiber Well, a fiber supplement in the shape of gummy candies, were chosen by 36% of the impersonators, compared to only 22% of the sample who were actually over 49. Furthermore, impersonators were less likely to choose the “none” option (7%) compared to the older respondents (25%). Both of these differences were statistically significant at a p<.05 level.

**Figure 2. Fiber Survey Choice Screenshot in Study 3**

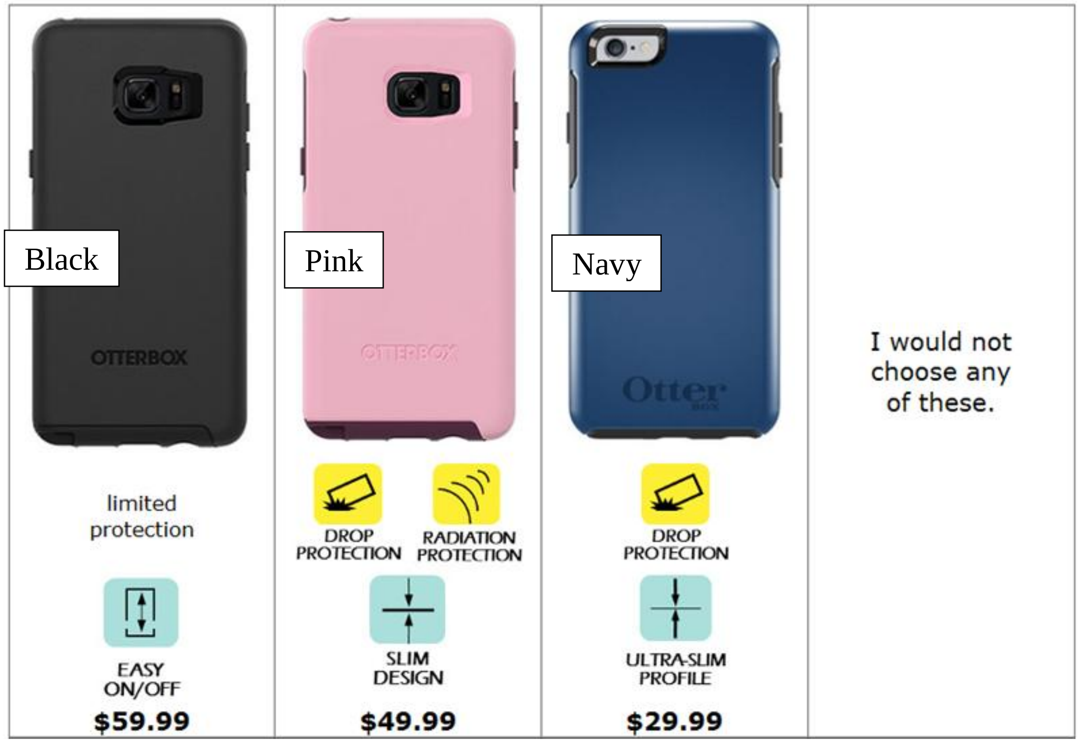


| | Actual Average Age | 90 Gummies<br>$14.99<br>○ | 100 Tablets<br>$15.99<br>○ | 125 Servings<br>$25.99<br>○ | I would not choose any of these products<br>○ |
|---|---|---|---|---|---|
| Honest about age | 56 | 22% | 50% | 3% | 25% |
| Dishonest about age | 33 | 36% | 51% | 7% | 7% |
| Re-Launch (no screener) | 34 | 47% | 36% | 3% | 15% |

To test whether some of the probable impersonators may have simply made a mistake in answering their age in the original panel survey or in the screening question, we ran an additional survey similar to the survey described above but did not screen out anyone (although a priori, we excluded those panelists who had already completed the screener version of the survey). When there was no monetary incentive to impersonate, ALL respondents provided the same age bracket as reported in the panel survey. This result is important, as it indicates that 40% of respondents claimed a different age when they could gain from it, but none did so when there was no motivation to lie. Interestingly, the respondents in the relaunch (average age of 34) chose the gummy candy fiber supplement 47% of the time, which was statistically different from the deceivers and those legitimately 50 years old or older.

The final test of the extent of character impersonation and its implications for market research results included a conjoint analysis study which examined gender impersonation. The announcement indicated that the 75 cent survey is only for females. Respondents were assigned randomly to either a control condition where there was no screener (but a gender question as part of the survey) or a treatment condition which included a gender screener. In the screener condition, males were disqualified unless they lied in the screener question about gender. Thus the control condition had males and females and the screener condition had males (pretending to be females) and females. In both conditions, respondents completed a 12 Choice-Based Conjoint (CBC) task based on cell phone case designs. The attributes (levels) for these cases, shown in Figure 3 included color (pink, black, or navy), style (slim design, ultra slim profile, or easy on/off of the case), drop protection (included or limited), radiation protection (included or limited), and price (ranging from $29.99 to $59.99).

**Figure 3. Example Choice Task from Conjoint Exercise in Study 4**



In the screener condition, 25% of the respondents claimed to be female but reported to be male in the panel survey. By contrast, in the control condition, all respondents reported the same gender in the conjoint study as in the panel survey. In analyzing the conjoint estimates, we found that males posing as females statistically differed from males in the control condition.

Specifically, males impersonating females tended to express their preferences in stereotypical ways by over-emphasizing their preference for attributes that they believed that females would like (i.e., color and design). They significantly preferred pink cell phone cases (average part-worth of 29.2) with an ultra-slim profile (average part-worth of 15.5) more than their male counterparts in the control condition (average part-worths of -127.1 and 2.5, respectively) and females overall (average part-worths of -9.0 and 6.0, respectively). For less stereotypical attributes such as drop and radiation protection, there was no statistical difference between males in the control condition and males posing as females in the screener condition. However, male and female utilities did significantly differ on these attributes, but this was not apparent to the deceivers. The same was true for the "none" option. Deceivers chose the "none" option just as often as their male counterparts in the control condition which tended, on average, to be more often than females.

As it relates to task consistency, the individual level "root likelihood" measure which is an output of the CBC Sawtooth Software, did not significantly differ between deceivers and non-deceivers. Similarly, the responses to the two fixed conjoint tasks (same choices, but different placement and order of the choices), showed no significant differences in consistency between the deceivers and non-deceivers. Deceivers did, however, spend slightly less time on the conjoint exercise and on an attitudinal scale statement section than non-deceivers, but did not statistically differ on the total time spent on the survey.

These results are important and potentially troubling for market researchers who utilize MTurk as a participant pool. We find in the fiber survey that impersonators project a youthful desire for fiber with gummy candies. In contrast, for cell phone cases, men who pretended to be women overcorrected in projecting women's desires for thin, pink phone cases. Worse, for other products, there was no distortion. Thus, it is very hard to know before the fact which way character impersonation will bias the results. These results, from a market research perspective, are simply unacceptable.

Now that we have repeated evidence of deception, we are able to observe across all four surveys if deception is an enduring trait (i.e., same people repeatedly deceive) or if most of our panelists are dishonest occasionally. Given the general finding in previous research that everyone (across a wide range of domains) basically cheats a little (Mazar, Amir, & Ariely 2008), we expected the latter. Our results, however, showed otherwise: about a third of our active panelists (35.8%) are consistently dishonest on 100% of the studies they took, about a quarter (25.2%) are dishonest on some of the studies, while the remaining panelists (38.9%) never deceive on screener questions. Thus we can conclude that a sizable proportion of our panel population will deceive some or all of the time. This makes our jobs easier (than if we found that everyone cheats a little all of the time), because we can exclude respondents who have evidence of deception from our panel in order to decrease the likelihood of deception in future studies.

Given we have the data about who consistently cheats, we are able to test if certain respondent characteristics are associated with cheating. In doing so, we found that persistent deceivers tend to be extroverted and male. Those who perceive themselves to be low on the socio-economic ladder are also more likely to deceive as well as those who perform poorly on the moral foundation sacredness scale (Graham & Haidt 2012). This latter scale measures one's tendency to engage in immoral behavior for money (e.g., how much would have to be paid to "kick a dog in the head, hard" with respondents stating how much they would have to be paid to do this or forgoing the money so as to not engage in the act). However, given the knowledge that certain characteristics are correlated with cheating, we do not suggest that researchers filter people out based on these attributes given one's likelihood of cheating on a study is not perfectly deterministic. Instead, filtering out respondents who consistently cheat (regardless of their respondent characteristics) is a better way to avoid future cheating behavior.

## Online Turker Communities

Given the significant proportion of our panel deceiving, we explored whether deception is acceptable amongst online Turker communities. These online forums tend to be Turker created and managed for the purpose of sharing advice about making money on MTurk. For example, TurkOpticon is a third-party site where Turkers rate researchers ("Requesters") on the speed of payment [FAST], compensation [PAY], communication [COMM], and fairness [FAIR]. These ratings along with qualitative comments serve as warnings (or approvals) of specific market researchers. Lower ratings can lead to Turkers being less willing to take future surveys from specific market researchers.

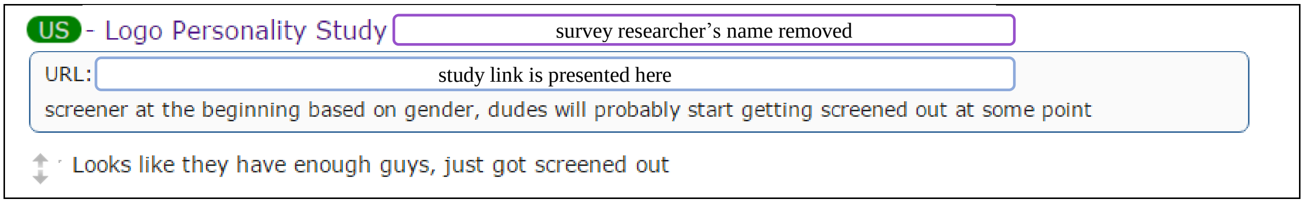
**Figure 4. TurkOpticon Example**

| | | |
|---|---|---|
| Requester Name*<br>A2____________M*<br>Averages »<br>HIT Group »<br>Review Requester » | FAIR: 1 / 5<br>FAST: 3 / 5<br>PAY: 1 / 5<br>COMM: 1 / 5 | Be careful! They check every single HIT and will easily reject you, if one single thing is wrong. Not worth the effort or time. Very low pay for rejects. Not worth it. |

* eliminated market researcher identification.

Another site, Hits Worth Turking For (HWTF), announces opportunities where the Turker can make at least 10 cents/minute based on the actual time taken rather than the estimated time posted by the market researcher. In MTurk terms, "Hits" ("Human Intelligence Tasks") are how someone refers to an MTurk task whether that task is a survey or some other activity (e.g., transcribing, categorizing websites, tagging photos, etc.). When posting on HWTF, fellow Turkers often warn their peers about the presence of a screening question (sometimes with the "correct" answer, see Figure 5 for an example) as well as any "trick" questions such as memory checks (MCs) or attention checks (ACs, see Figure 6 for an example).

**Figure 5. Example of a Screening Question Warning on Hits Worth Turking For**



US - Logo Personality Study survey researcher's name removed

URL: study link is presented here

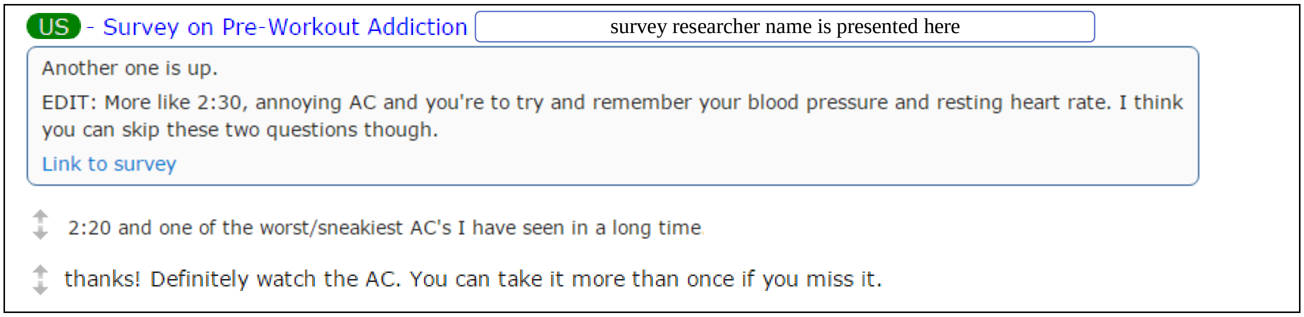
screener at the beginning based on gender, dudes will probably start getting screened out at some point

Looks like they have enough guys, just got screened out

Source: Reddit Forum: Hits Worth Turking For

In general, memory checks are used by market researchers who are interested in catching if a person is paying attention to information provided earlier in the survey. Attention checks often include tricky wording or encourage impossible answers to those who are not paying sufficient attention to the survey. Respondents can be dropped from a survey for failing such tests, thus the motivation for Turkers to warn their peers. In general, experienced Turkers, with more than 1,000 completed tasks, know to be on alert. We find that our panelists, who are largely experienced users, are relatively immune to the standard attention and memory checks. Indeed, those who impersonate to get into a survey are generally careful respondents. In the conjoint cell phone case study for example, those who impersonate were not less or more likely to fail the included memory check or attention check (adapted from Goodman, Cryder, & Cheema 2013). The point here is that it is hard to identify qualification cheats from patterns or inconsistency of their responses per traditional quality measures (e.g., time, attention checks, memory checks, etc.). They are in a performance sense, ironically good respondents.

**Figure 6. Attention Check (AC) Warning Example from Hits Worth Turking For**



US - Survey on Pre-Workout Addiction

survey researcher name is presented here

Another one is up.

EDIT: More like 2:30, annoying AC and you're to try and remember your blood pressure and resting heart rate. I think you can skip these two questions though.

Link to survey

2:20 and one of the worst/sneakiest AC's I have seen in a long time

thanks! Definitely watch the AC. You can take it more than once if you miss it.

Source: Reddit Forum: Hits Worth Turking For

It was in observing this online behavior that caused us to wonder to what extent these online forums promote dishonesty. While we have personally observed numerous examples where warnings could lead to deception, we were interested in understanding if the repeat offenders in our panel were more likely to be active on one or more of these sites. Thus as part of our panel survey, we had given participants the option to self-disclose the communities in which they participate (see the list in Table 1). Of our panel respondents, 66% mentioned that they participated in at least one of these MTurk-related online communities.

**Table 1. Online Turker Communities**

| Name<br>(website) | Purpose | proportion of panelist* |
|---|---|---|
| Hits Worth Turking For (HWTF)<br>(https://www.reddit.com/r/HITsWorthTurkingFor) | To notify Turkers about tasks that pay at least 10 cents/minute and warn fellow Turkers about tasks which include screening questions, attention checks (AC), and/or memory checks (MC). | 34% |
| MTurk Grind<br>(http://www.mturkgrind.com/) | To help Turkers be successful on MTurk. Additionally, much of the chatroom discussion involves venting about Turking or daily life in general. It provides its users with a sense of community. It is also the place where scripts (browser tools that make finding good HITs easier and faster) are announced. | 23% |
| Turk Opticon (TO)<br>(https://turkopticon.ucsd.edu/) | To rate “Requesters” (i.e., market researchers or anyone requesting “work” to be done by a Turker) in regard to pay, fairness, speed, and communicability. | 20% |
| MTurk Forum<br>(http://www.mturkforum.com/) | To give advice and brag about one’s success (e.g., money made, HITs completed, etc.) on MTurk. | 17% |
| Turker Nation<br>(http://www.turkernation.com/) | To “benefit the [Mturk] workers—allowing freedom of discussion of HITs, rating of requesters, talking about how to make more money, etc.” Registration is required to read or participate and Requesters are permitted to join, but with limited access to forums. | 9% |

* based on optional self-report in our panel survey. Participants can be a part of multiple communities.

Given this information, we compared those who consistently deceive with those who do not to see if impersonators were more likely to participate in one of these online communities. We were surprised to find that those who engaged in these sites, were actually less likely to be dishonest. Admittedly, it is possible that those who deceive are less likely to report frequenting these sites which would bias our results. Still, in our panel surveys, we have little evidence that participants deceive (or hide information) when there is no incentive to do so. Regardless, we recommend that market researchers monitor in real time these sites when a survey is live in case there is any information shared amongst respondents which could potentially harm the data.

That said, we do not believe that these online communities are the primary source of deception, but that instead, deception is an inconspicuous action taken by a subsection of individual Turkers (e.g., by clearing cookies in a browser and making multiple attempts at a screener question on a survey). Indeed, Turker communities largely serve to help fellow workers identify jobs that will not be too onerous and will provide a reasonable economic return for work from home.

Accordingly, we recommend that market researchers develop and test their own panel of Turkers as we did and remove any respondents who prove to be problematic over time by repeatedly deceiving. Doing so allows researchers to pre-qualify respondents based on certain characteristics (identified in the panel survey); thus, only make a survey available to those who would otherwise qualify. Like professional panel companies, this is done by keeping a database of respondents and flagging suspicious activity. In this way, researchers can benefit from the inexpensive participant pool that MTurk provides while limiting exposure to deception.

## In Brief

Not unlike other online participant pools, Turkers tend to be substantially younger than the general US population and have slightly lower household incomes. From this participant pool, we found that deception rates (lying about their qualifications in screener questions) to get into studies ranged from 25%–82% which is likely a function of how difficult it is to qualify for the screener (i.e., the harder it is to qualify, the higher the rate of cheating). While deception can occur through online user sites, it appears that active deception is more inconspicuous. Unsurprisingly, Turkers will not lie when there is no benefit of doing so (like in our original panel survey where respondents did not know that we were going to use the information for future surveys). They are conscientious and consistent when there is no motive to deceive.

For those interested in using MTurk as a relatively inexpensive participant pool, we recommend building your own panel. It is easy and relatively cheap to build this panel, but must be tested for consistency so respondents who deceive can be quietly removed. An added benefit of this panel is that it would serve well for those needing a longitudinal sample. If creating a panel is not feasible, screening on MTurk should be done outside of the primary survey (in a pre-survey) without announcing that it is a screener.

Kathryn Wessling

Joel Huber

Oded Netzer

## REFERENCES

Chandler, J., Mueller, P. & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers. Behavior Research Methods, 46 (1), 112–130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P. & Ratliff, K. A. (2015). Using Nonnaïve Participants Can Reduce Effect Sizes. Psychological Science, 26 (7), 1131–1139.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). “Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples.” Journal of Behavioral Decision Making 26 (3), 213–224.

Graham, J., & Haidt, J. (2012). Sacred Values and Evil Adversaries: A Moral Foundations Approach. The Social Psychology of Morality: Exploring the Causes of Good and Evil, 11–31.

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. Journal of Marketing Research, 45 (6), 633–644.

Stewart, N., Ungemach, C., Harris, A., Bartels, D., Newell, B., Paolacci, G. & Chandler, J. (2015). The Average Laboratory Samples a Population of 7,300 Amazon Mechanical Turk Workers. Judgment and Decision Making, 10 (5), 479–491.

Weinberg, J. D., Freese, J. & McElhattan D. (2014). “Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-based and a Crowdsource-recruited Sample.” Sociological Science, 1, 292–310.

# PROCESS TRACING: A NEW TOOL FOR MODELING PHYSICIAN TREATMENT ALGORITHMS

**_STEPHEN BELL_**
**_DOUGLAS WILLSON_**
*A+A BELL FALLA*

In recent years, the forces that impact healthcare decision-making have come under increasing scrutiny. Advances in clinical investigation, data analysis, and rapid dissemination of information have led to the development of standardized guidelines and treatment algorithms in many therapeutic categories. At the same time, physicians are under ever-increasing time pressure and are flooded with daunting amounts of new information on a daily basis. HCP diagnostic and treatment decisions are one of the most important influences on patient outcomes. New technologies and electronic medical records support evidence-based medicine and use of standardized algorithms for both diagnosis and treatment decisions—yet evidence suggests that physicians regularly deviate from published guidelines in some therapeutic categories, relying on personal experience and heuristics to make decisions (Groopman 2008). From a marketing perspective, research on HCP decision-making provides important insight on the drivers of market share for pharmaceutical brands. From a public policy perspective, research on HCP decision-making may provide clues to misdiagnosis and suboptimal treatment recommendations in challenging therapeutic categories.

In this paper, we discuss a new market research tool—process tracing—that can be used to identify and investigate physician treatment algorithms in today's rapidly changing pharmaceutical markets. Process tracing is a method for investigating how physicians and other healthcare providers make diagnosis and treatment decisions. Originally developed for consumer market research on choice processes conducted in the 1970's, process tracing investigates the steps physicians take to acquire information about their patients, as well as their ultimate diagnosis or treatment decisions. The process tracing approach can be conveniently implemented in an online survey environment as a relatively simple modification of a traditional choice exercise.

To illustrate the application of these methods, we provide a detailed case study involving a survey of 200 physicians in the US and EU. We provide details on survey design, model specification, and post-survey analysis. In the case study, we show how raw data from the survey provides information regarding the choice process that is not available with the traditional choice modeling methods. In particular, the process tracing exercise provides detailed information on the depth of information search—how much information is required to make a treatment decision—as well as the sequence of patient attributes investigated (a proxy for attribute importance). The new choice model can be used to summarize an algorithm for the market as a whole, or to develop a physician segmentation based on the algorithms they use for treating patients. Results from the process tracing approach are also compared and contrasted with a traditional discrete choice modeling approach.

Our paper begins with a brief introduction to process tracing and the information board, comparing and contrasting this approach with other methodologies. Section 2 describes the

PT/IB choice model. Section 3 presents the case study. Section 4 concludes and makes some suggestions for future research.

## BACKGROUND—PROCESS TRACING AND THE INFORMATION BOARD

Stedman's Medical Dictionary defines an algorithm as . . .

> *A systematic process consisting of an ordered sequence of steps, each step depending on the outcome of the previous one. In clinical medicine, a step-by-step protocol for management of a health care problem.*

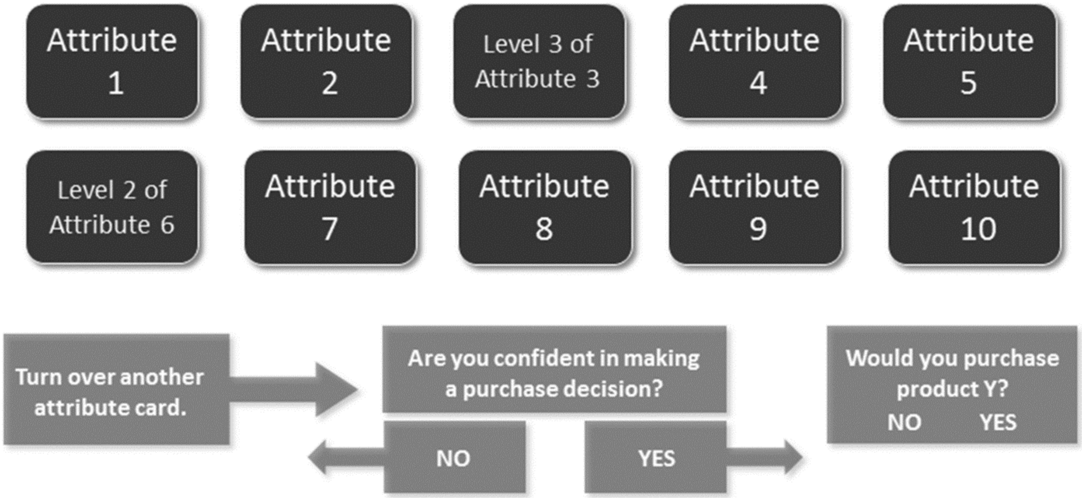
In pharmaceutical markets, there are many reasons to suggest that healthcare professionals use algorithms to make diagnosis and treatment decisions. Perhaps most importantly, regulatory authorities often produce guidelines for managing patient care. Guidelines are systematically developed statements designed to assist practitioners and patients in making decisions about appropriate healthcare for specific clinical situations. They typically represent a consensus on the best information medical science has to offer in specific clinical situations, and in many situations these guidelines are delivered in an algorithmic format. This behavior is reinforced by payers who often require that physicians follow guidelines for reimbursement of medicines and healthcare services.

Traditional preference models that are regularly used in marketing research assume that individuals evaluate and weigh/trade-off attributes to make a choice. Discrete choice experiments and stated preference methods using this framework are now commonplace, clearly evidenced by the large number of papers presented and referenced at Sawtooth Software conferences over many years. There is also an established literature considering alternative non-compensatory theories of choice—models with screening rules (Gilbride and Allenby 2004, 2005), lexicographic preferences (Kohli and Jedidi 2007), and other modifications have also been investigated. Many of these alternative approaches can be viewed as heuristics—short, simplified decision rules that may be optimal in some contexts.

A related literature investigates theories of choice by asking market participants to report on how and why they make their decisions while they are making them—these process tracing (PT) methods have been used in marketing research to investigate information processing in choice experiments (Lohse and Johnson 1996, Zhu and Timmermans 2010). More generally, they have been utilized in many research contexts to investigate the "hows" and "whys" of decision-making; these methodologies are summarized in the Handbook of Process Tracing Methods for Decision-Making (Schulte-Mecklenbeck *et al.* 2010).

Information boards (IB) were originally developed for investigating choice processes and measuring attribute importance in consumer research (Payne 1976). In a typical application, the names of product attributes were placed on cards arranged on a large board. Consumers were asked to turn over the cards sequentially to reveal underlying product attributes and to subsequently make a purchase decision when enough information was displayed. Information boards provide detail on the depth and direction of information search strategies, as well as the impact of specific information on decisions. The PT/IB choice task is described in Figure 1 below, and can be easily implemented in an online survey.

**Figure 1. PT/IB Choice Tasks**



Not surprisingly, IB methods have also been used to investigate healthcare treatment and diagnosis decisions as part of the process tracing methodology (Chinburapa *et al.* 1993). In healthcare applications, HCPs turn over cards to reveal specific patient characteristics, and are then asked whether they are confident in making a prescribing or diagnosis decision, or would like to see more patient information.

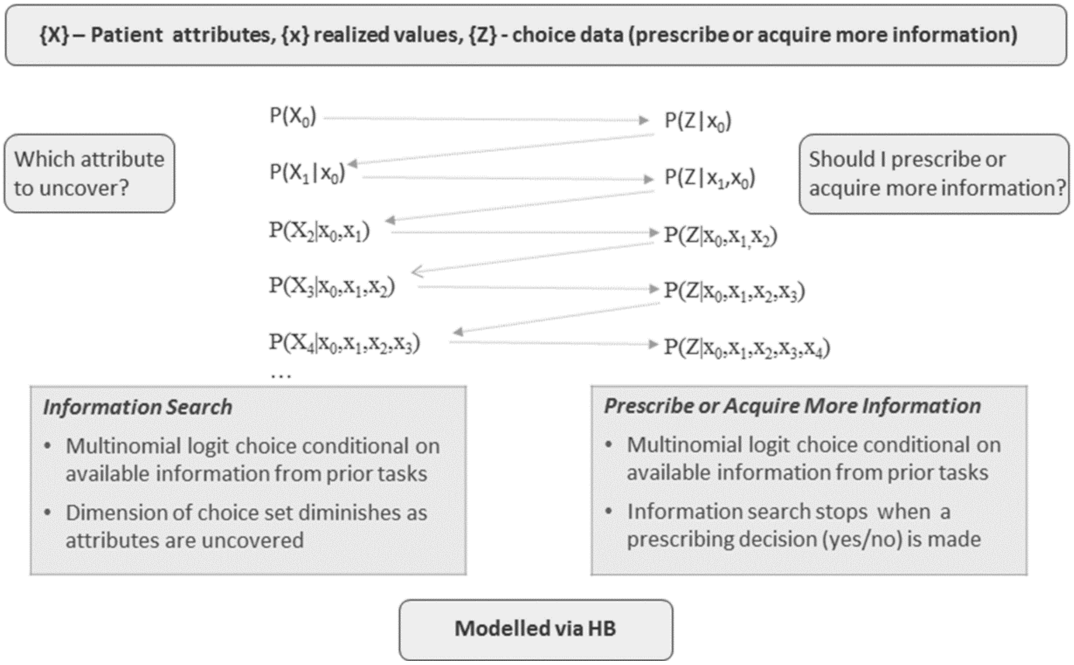
## MODELING WITH PT/IB DATA

The PT/IB model simultaneously investigates two choices:

1. Information search—which patient characteristic is uncovered next?
2. Should I prescribe or acquire more information?

The model structure is described in Figure 2 below:

**Figure 2. PT/IB Model Structure**


{X} – Patient attributes, {x} realized values, {Z} - choice data (prescribe or acquire more information)
Which attribute to uncover?
$P(X_0)$
$P(X_1|x_0)$
$P(X_2|x_0,x_1)$
$P(X_3|x_0,x_1,x_2)$
$P(X_4|x_0,x_1,x_2,x_3)$
…
$P(Z|x_0)$
$P(Z|x_1,x_0)$
$P(Z|x_0,x_1,x_2)$
$P(Z|x_0,x_1,x_2,x_3)$
$P(Z|x_0,x_1,x_2,x_3,x_4)$
Should I prescribe or acquire more information?
Information Search
Multinomial logit choice conditional on available information from prior tasks
Dimension of choice set diminishes as attributes are uncovered
Prescribe or Acquire More Information
Multinomial logit choice conditional on available information from prior tasks
Information search stops when a prescribing decision (yes/no) is made
Modelled via HB

It is possible to estimate the model components jointly as sequential multinomial logit choices via HB. It is also possible, under certain assumptions, to model each component separately via HB MNL. Generating a simulation from the model is not trivial. Since we are investigating both the order in which attributes are investigated as well as the impact of revealed attribute levels on choices, with even moderate numbers of patient attributes there are billions of possible algorithms to consider. However, as with standard decision trees, relatively simple heuristics can be used to generate algorithms that are useful. For the case study we describe in the next section, we use a simple first choice heuristic to develop algorithms. Research for additional approaches to quickly generate algorithms for these models is ongoing.

## Case Study

As an illustration of the PT/IB approach, we present a case study investigating the impact of patient attributes on prescribing for a new drug in development to accelerate fracture healing. The study involved an online survey of 200 orthopedic surgeons in the US and EU, and included a novel application of the information board and a comparison of results with traditional discrete choice methods.

More specifically, the study considered patients with tibia fractures. The tibia (or shinbone) is the larger of the two bones in the leg below the knee, and is the most frequently fractured "long" bone in the body (long bones include the femur, humerus, tibia, and fibula). The tibia has poor blood supply because it is surrounded mostly by skin and fat, instead of muscle; poor blood supply inhibits fracture healing. Typical healing problems include malunion (the bone heals in the wrong place), nonunion (the fracture never heals), and delayed union (healing takes longer than expected). Many clinical trials have investigated the use of osteoporosis drugs (e.g., bisphosphonates, parathyroid hormone [PTH] therapy) to accelerate fracture healing, but results have not been successful to date. In this therapeutic area, the PT/IB approach was viewed as potentially quite useful because there is no standard clinical definition for characterizing a healed fracture (Morshed *et al.* 2008), there are no benchmark pharmaceutical standards of care, and approaches to treatment with the new therapy could vary widely across physicians depending on their evaluation of patient characteristics and interest in the new therapy.

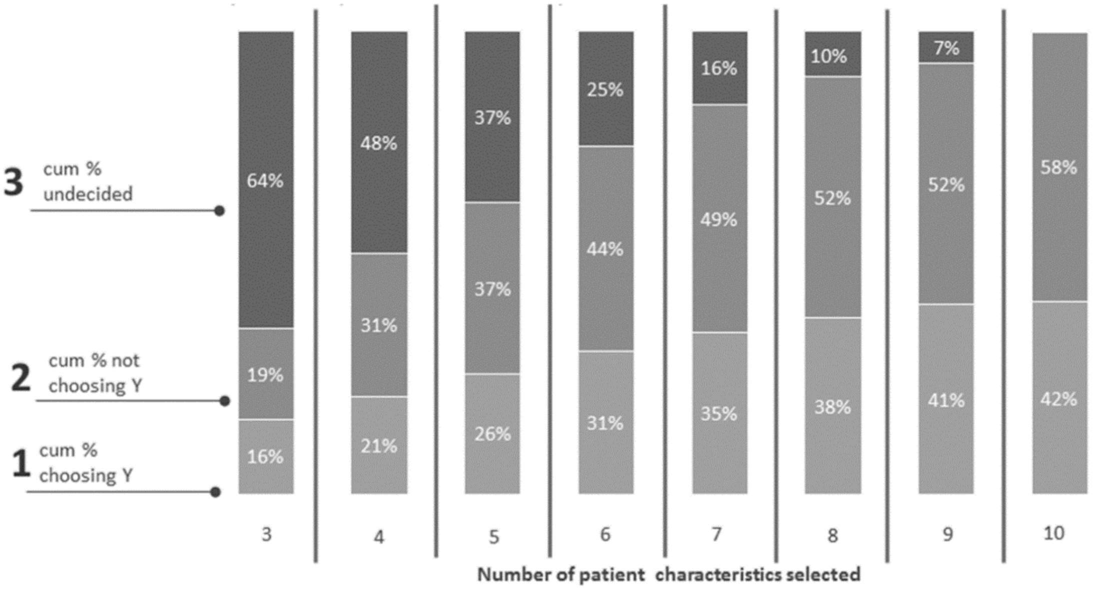
For the PT/IB exercises, patients were defined using patient attributes that included fracture characteristics, patient background variables, and other characteristics, such as length of hospital stay. Hypothetical patient cases were generated by experimental design. Physicians evaluated 8 different patient cases in the PT/IB choice tasks. For each case, physicians were instructed to select patient attributes in order of their importance for making a treatment decision regarding use of the new product, Product Y. After selecting a patient characteristic, physicians were asked: "Are you . . ."

- Confident in prescribing Product Y for this patient.
- Confident in NOT prescribing Product Y for this patient.
- Undecided, would like to see more patient characteristics.

Physicians continued selecting patient attributes until they were comfortable making a prescribing decision (Yes/No) for Product Y. If a physician continued without making a choice through 10 patient attributes, they were asked to make a final decision at that time.

Figure 3 below shows the raw choice data summarized across all PT/IB tasks. Patient cases in Figure 3 are identified by the number of patient attributes selected when physicians made their definitive prescribing decision. Across all choice tasks, Figure 3 suggests that physicians made a prescribing decision in 36% of patient cases with 3 attributes selected; allowing physicians to select up to 9 attributes resolved the majority of uncertainty, with prescribing decisions made for 93% of patient cases.

**Figure 3. Raw Choice Data Summary by Number of Patient Characteristics Selected**



Figures 4a and 4b provide additional information on the choice exercises. While physicians evaluated 8 different patient cases, the experimental design was blocked, so different physicians saw some of the same patient cases. The average number of patient characteristics selected for each fixed case in the design is shown in Figure 4a. Figure 4b provides a frequency distribution for the average number of characteristics selected per physician.

**Figure 4a**
**Average Number of Patient Characteristics Selected Across Fixed Patient Cases**

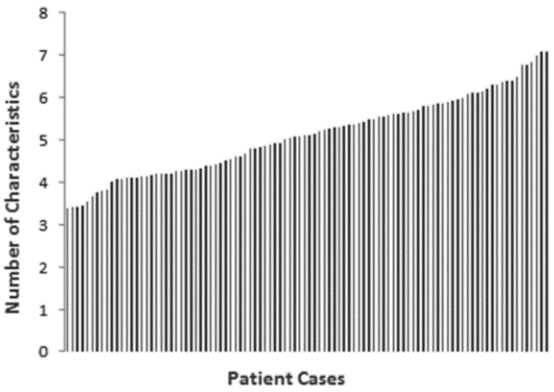
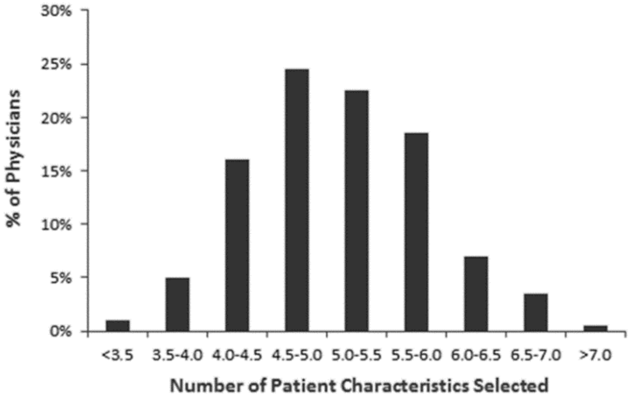


**Figure 4b**
**Average Number of Patient Characteristics Selected Per Physician**

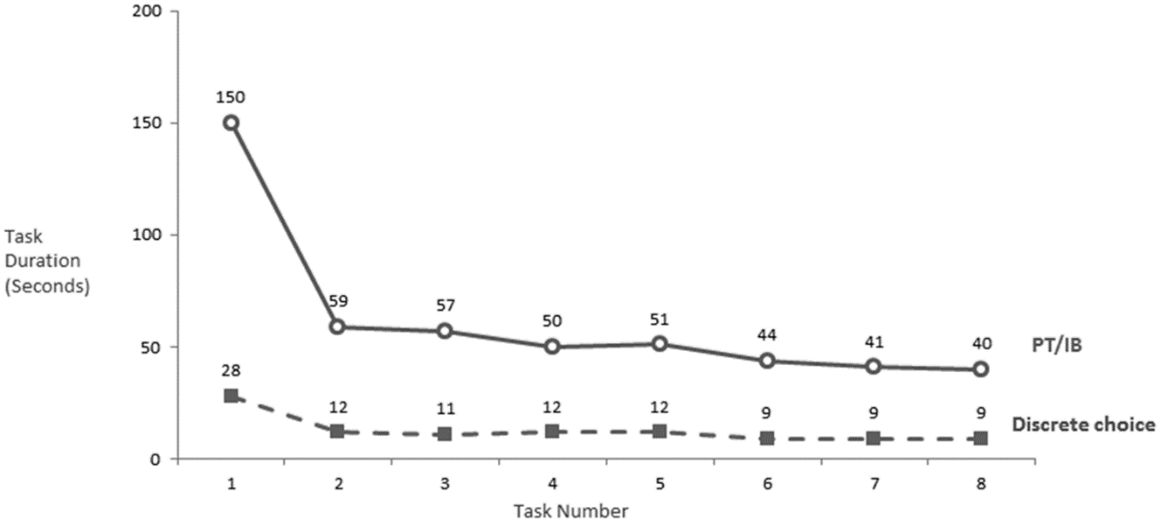


These figures illustrate that the depth of information search varies by physician and across patients. Not surprisingly, some physicians make decisions more quickly than others, and some patients are more “difficult” to treat than others—so more information is required to make a treatment decision.

The case study also included a comparison of the PT/IB approach with traditional discrete choice exercises. After completing the 8 PT/IB patient cases, physicians were asked to complete another 8 discrete choice exercises with different patient cases. In the DC exercises, all patient attributes were displayed and the physicians were simply asked whether or not they would prescribe Product Y for each patient.

Figure 5 displays the average survey task durations for the PT/IB and DC exercises. For both methodologies, the initial tasks take the most time, and reductions in task duration occur as respondents complete more exercises.

**Figure 5. PT/IB and DC Task Durations**



The initial PT/IB task duration—a whopping 150 seconds—stems from an extensive set of instructions provided to respondents to ensure they clearly understood the PT/IB task in the survey. In future research we plan to investigate if this can be shortened to some degree. A comparison across methodologies suggests the PT/IB tasks take 4 to 5 times as long as traditional DC tasks—the difference reflects the extra time associated with collecting the process information and the sequence of attributes investigated for each patient case in the PT/IB approach.

With traditional DC exercises, reductions in choice task duration as the number of tasks increase have been associated with learning and task simplification in the survey environment (Allenby *et al.* 2005, Johnson and Orme 1996). We also investigated the number of patient attributes selected in the PT/IB approach as respondents completed more exercises. While PT/IB task duration decreases with the number of exercises, the average number of attributes selected hovered around 6 across all exercises. At least for the PT/IB approach, these results suggest that it is not task simplification per se (i.e., looking at or concentrating on fewer attributes) so much as learning about the exercises within the survey environment that produces the reduction in task duration as the survey progresses.

We also compared results from the PT/IB exercises and modeled choices with results from the traditional DC choice exercises. Figure 6a reports the percentage of patient cases where Product Y was prescribed for the two approaches. While individual respondents evaluated different patient cases for separate sets of exercises, the designs were balanced across exercises, so we should expect roughly the same rates of prescribing under each approach if PT/IB truly captures all of the relevant information for each patient. Figure 6a suggests overall prescribing rates are very similar across methodologies.

**Figure 6a**
**% of Patient Cases Where Product Y is Prescribed**

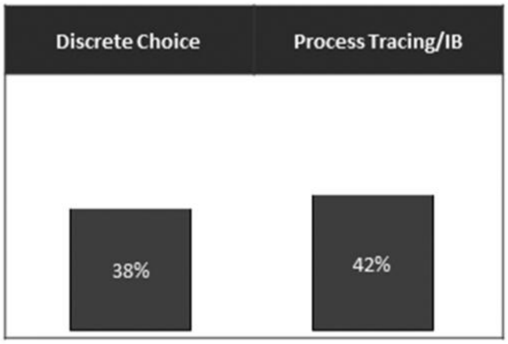


**Figure 6b**
**% of Out-of-Sample Patient Cases Where Modeled Selection = Hold Out Selection**

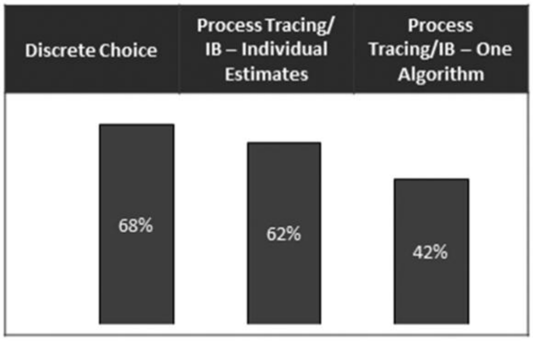


Figure 6b compares out-of-sample forecasting performance for PT/IB versus DC. For this analysis, we removed one of the choice tasks at random from the choice sets for each approach for each respondent, estimated the model with the remaining 7 exercises, and then generated a first choice forecast based on the estimated models and the held out patient cases. The DC choice model is estimated via HB using a standard binary logit. To generate a forecast for the PT/IB model, we use a simple first choice heuristic.

1. Select the attribute with highest probability of being selected in model 1.
2. For each patient case in the sample, compute prescribing probabilities for Model 2, conditional on the levels of revealed attributes. In each terminal node, if P (Yes or No) (i.e., the probability of making a decision) >P, some minimum threshold, stop.
3. Otherwise, continue and, select the next attribute with the highest probability of being selected in Model 1. Repeat step 2.

After the heuristic has stopped for every patient case, we compute the predicted choice for each case as the highest predicted probability (Yes or No) and compare it with response from the holdout task. These results are displayed in Figure 6b.

In Figure 6b, results for the traditional discrete choice approach are slightly better than PT/IB with individual estimates. The PT/IB model where all physicians use the same algorithm fares poorly—heterogeneity is clearly a characteristic of physician prescribing in this market, and one should not expect an aggregate model to perform well in this situations.

We also compared traditional estimates of attribute importance from the DC exercises/model with those from the PT/IB approach. For the latter, we estimated attribute importance using the percentage of times the attribute was selected across all PT/IB tasks. Table 1 displays the importance rank (1st is more important) for each methodology.

**Table 1. Comparison of Attribute Importance Estimates**

| | | Attribute Importance Rank | |
|---|---|---|---|
| Category | Attribute | PT/IB | DC |
| Fracture Characteristics | FC1 | 4 | 5 |
| | FC2 | 7 | 8 |
| | FC3 | 2 | 1 |
| | FC4 | 3 | 3 |
| | FC5 | 6 | 7 |
| | FC6 | 5 | 4 |
| Patient Background | PB1 | 9 | 6 |
| | PB2 | 11 | 12 |
| | PB3 | 8 | 10 |
| | PB4 | 10 | 13 |
| | PB5 | 13 | 12 |
| | PB6 | 12 | 11 |
| Other | O1 | 1 | 2 |

Overall, the attribute importance ranks appear to be quite consistent when averaged across respondents and patient cases.

## CONCLUSIONS

Process tracing is a method for investigating how physicians and other healthcare providers make diagnosis and treatment decisions. Originally developed for consumer market research on choice processes conducted in the 1970's, process tracing investigates the steps physicians take to acquire information about their patients, as well as their ultimate diagnosis or treatment decisions.

The process tracing approach can be conveniently implemented in an online survey environment using the information board, a relatively simple modification of a traditional choice exercise.

Research suggests that the PT/IB approach provides new information concerning:

- Patient attribute importance
- Patterns and depth of physician information search
- Stopping rules
- Physician treatment algorithms

Results of process tracing modeling with heterogeneity compare favorably with traditional discrete choice results. Future research will further investigate modeling approaches and different heuristics for generating algorithms and predictions for the PT/IB approach.

Stephen Bell

Douglas Willson

## References

Allenby, G., *et al.* (2005) Adjusting choice models to better predict market behavior. Marketing Letters, 16:3/4, 197–208.

Baiardini. I., Braido, F., Bonini, M., Compalati, E., and G. Canoninca (2009) Why do doctors and patients not follow guidelines. Current Opinion in Allergy and Clinical Immunology, June; 9(3) 228–233.

Chinburapa, V. *et al.*, (1993) Physician prescribing decisions: the effects of situational involvement and task complexity on information acquisition and decision making. Soc Sci Med 36, 1473–1482.

Gilbride, T., and G. Allenby (2004) A choice model with conjunctive, disjunctive, and compensatory screening rules. Marketing Science 23(3), 391–406.

Gilbride, T., and G. Allenby (2005) Estimating heterogeneous EBA and economic screening rule choice models. Marketing Science 25(5), 494–509.

Groopman, J., (2008) How Doctors Think. Mariner Books.

Jacoby, J. (1977) “The emerging behavioral process technology in consumer decision-making research,” in Advances in Consumer Research Volume 04, eds. William D. Perreault, Jr., Atlanta, GA: Association for Consumer Research, 263–265.

Johnson, R., and B. Orme (1996) How many questions should you ask in choice-based conjoint studies. Technical Report, Sawtooth Software.

Kohli, R., & Jedidi, K. (2007). Representation and inference of lexicographic preference models and their variants. Marketing Science, 26, 380–399.

Lohse, G., and E. Johnson (1996) A comparison of two process tracing methods for choice tasks. Organizational Behavior and Human Decision Processes, Vol. 68, No 1, October, 28–43.

Morshed, S., Corrales, L, Genant, H., and T. Miclau (2008) Outcome assessment in clinical trials of fracture healing. Journal of Bone and Joint Surgery, Feb; 90 Suppl 1: 62–67.

Payne, J (1976) Heuristic search processes in decision-making. in Advances in Consumer Research Volume 03, eds. Beverlee B. Anderson, Cincinnati, OH : Association for Consumer Research, 321–327.

Schulte-Mecklenbeck, M, Kuhlberger, A., and R. Ranyard, eds. (2010) A Handbook of Process Tracing Methods for Decision Research. Society for Judgement and Decision Making, Psychology Press.

Zhu, W, and H. Timmermans (2010) Modeling simplifying information processing strategies in conjoint experiments. Transportation research Part B 44, 764–780.

# LET'S TAKE NONE SERIOUSLY

***ULA JONES***
***TOMER J. OZARI***
***PETER KURZ***
*TNS*

## ABSTRACT

Choice deferral, the "none" alternative in a discrete choice task, has traditionally been asked in a single stage free choice task, where "none" is another alternative, or in a dual-response forced choice task, where selection is followed by evaluation as a way to estimate the "none" utility. Some researchers treat the "none" alternative only as a way of allowing respondents to opt out of a difficult choice scenario while other researchers also view the "none" alternative as a means of estimating likelihood not to purchase, thus, gaining insight into product or service demand. Since it is not always possible to calibrate choice deferral with secondary data, as in the case of new products or services, we further examine the outcome regarding different approaches of asking the "none" option. While some researchers favor the traditional "none" over the dual-response approach, it is argued that the dual-response approach, where selection precedes evaluation, is predisposed to reduction of cognitive dissonance and a reversed dual response, where evaluation precedes selection, is proposed as an alternative.

## INTRODUCTION

The traditional rating based conjoint (Green & Rao 1971), where a single product or service profile is rated in terms of willingness to purchase, was a good beginning to the tradeoff task but not ideal since a rating is further removed from a behavior and an estimation of respondents who would not prescribe to the offering could not be derived. The work of McFadden (1974), utilizing the conditional logit as an econometric model of population choice behavior based on the decision rules of individuals, was later combined with an experimental design within a multinomial logit framework in choice modeling by Louviere and Woodworth (1983), commonly known as discrete choice analysis. The introduction of discrete choice modeling was an improvement to the rating based conjoint because a discrete choice was closer to a behavior, a behavioral intent rather than a rating.

In discrete choice analysis the "none" option was offered as another alternative, thus providing the benefit of the respondent to opt out of making a choice as well as allowing for an estimation of the proportion of respondents who would not be interested in the current product line. But the value of the traditional "none" option is debatable because selecting the "none" alternative during the choice task could be an escape strategy to make the exercise less burdensome but it can also serve to balance the utilities in situations where alternatives are either high or low on features, making the decision more difficult. It has been suggested to estimate the "none" parameter but not to include the "none" alternative in market estimations because it is not clear why respondents choose "none" during the discrete choice task (Sawtooth Software 2013). It is therefore suggested that a better understanding of why and when respondents choose the "none" alternative during the tradeoff task will make the "none" parameter useful during simulations as a way to estimate choice deferral.

According to Karty and Yu (2012), the estimation of the “none” option as a means of determining the proportion of respondents not interested in the combination of offerings could be biased, in most cases overestimating the likelihood of purchase. Researchers are often interested in the estimation of the “none” utility in a discrete choice model as a means of estimating share of preference for those not interested in the offering. Any bias in the deferred choice response (the “none” option) will cause the model to either over-predict or under-predict the likelihood to purchase the product or line of products. It is often the case that the “none” option has traditionally been under-selected, thus, overestimating willingness to purchase (Breidert, Hahsler & Reutterer 2006, Karty & Yu 2012).

Overestimation of willingness to purchase when “none” is another alternative in the choice set has evolved into the dual-response approach (Uldry, Diener & Severin 2002, as cited in Brazell, Diener, Karniouchina, Moore, Séverin & Uldry 2006). The dual-response technique initially presents the respondent with a forced choice of alternatives, without the “none” option. Upon making a selection, the respondent is then presented with a follow-up question regarding willingness to purchase the product just chosen. This follow up question allows for the estimation of the “none” utility in the multinomial logit model. The dual response increased the proportion of “none” over the traditional single-stage technique where “none” was another alternative in the choice set. In addition, the dual response provided more data for estimating the feature levels than the single-stage approach, especially in cases where the “none” alternative was overly chosen. Although the dual response has increased the choice share for “none” and allowed for more data to estimate the feature levels, there was still the plaguing issue of overestimating willingness to purchase when compared to real-world data.

Differences in the proportion of choice deferral as a function of either the traditional “none” or dual-response methodology is further confounded by the number of attributes in the study. Choice deferral is treated differently in studies with two attributes, such as product and price, versus studies with a larger number of attributes. Tradeoff studies with fewer attributes often have a clear compromise alternative, whereas studies with more attributes do not have clear compromise alternatives. The lack of a clear compromise alternative in multi-attribute studies suggests that “none” has a lower disproportionality in drawing preference from the other alternatives as compared to the presence of “none” in a study with fewer attributes (Dhar & Simonson 2003). In studies where the focus is on fewer attributes it is assumed that there is a larger disproportionality when adding or deleting the “none” option due to a clear compromise option in the choice tasks.

To reduce the likelihood of purchase that is often an overestimation in survey research, the dual response was converted from a binary yes/no scale of purchase intent to a 5-point scale, where the researcher was able to further calibrate willingness to purchase by considering top box as purchasing the selected product and the remainder of the scale as indication of non-purchase (Karty & Yu 2012). However, manipulating the cutoff of purchase intent is subjective and does not address the underlying issue of measurement validity of choice deferral.

Acknowledging that stated choices (from a choice task) are different from revealed preferences (real-world behavioral data), another option to fine tuning the “none” option in a choice task (either the traditional single-stage none or dual-response methodology) has been proposed by Ben-Akiva, Bradley, Morikawa, Benjamin, Novak, Oppewal, & Rao (1994) through combining revealed and preference data. Combining multiple sources of data (revealed preference, stated intention, stated choice, stated judgment ratings, attribute ratings, similarity ratings, attitudinal ratings, background characteristics of the respondents) allows for a more accurate estimation of choice models for both the features as well as the “none” alternative.

The unifying framework of Ben-Akiva *et al.* (1994) is a reasonable approach to correct for biased survey responses by calibrating all model parameters, including the estimation of the “none” utility. However, this modeling framework that increases the external validity of the predictive models has limitations. According to Ben-Akiva *et al.* (1994):

> *[I]t is not possible to use revealed preference data to estimate alternative-specific variables in the case of a new alternative that does not currently exist in the marketplace. Since predicting demand for new alternatives is one of the major reasons for collecting stated preference data, this situation arises quite often. In such cases, we have no choice but to rely on stated preference data for attributes or constants specific to the new alternatives. (p. 347)*

Given that the “none” option does not draw proportionately from all alternatives, especially when the new alternatives do not exist in the marketplace, the unifying framework of combining revealed and stated preference data to calibrate the “none” alternative is not a viable option in all cases, suggesting the need to further examine the most ecologically valid approach towards measuring choice deferral.

Choosing the “none” alternative is not only a function of how it is asked, either as another alternative in a free-choice task (traditional “none”) or after making a selection in a forced-choice task (dual response), it is also influenced by the alternatives offered in the competitive set. Tversky and Shafir (1992) have examined the implication of choice under conflict on deferred decision, the decision to choose the “none” option. Students were presented with two apartment options varying in distance from campus and monthly rent. They were given the option of choosing either apartment or continuing to look for apartments at the risk of losing one or both of the apartments. The results show that choice deferral in search of additional alternatives depends not only on the best available options but also in the difficulty of choosing among the existing options when they are equally attractive. When the selection decision among the competitive set of alternatives is difficult because there are at least two attractive options people would rather defer their decision as a way to reduce their current state of conflict.

On the basis of this reasoning, that choosing the “none” alternative is a choice deferral strategy to reduce the state of conflict, we suggest that “none” will be chosen less often when it proves not to be a choice deferral strategy. We present respondents with the traditional “none” discrete choice task, however, in scenarios where “none” is selected, the respondent is then presented with a forced choice. Forcing the respondent to choose a product as a follow-up when the “none” alternative is selected suggests that conflict is not reduced because product selection is still required. Hence, we propose the first hypothesis:

**H1**: Traditional “none” (a free choice task with “none” as another alternative) with a follow up forced-choice when the “none” alternative is chosen will have a lower proportion of choice deferral than the traditional “none.”

Furthering the research on decision making, Dhar and Simonson (2003) introduced Leon Festinger’s (1964) theory of Cognitive Dissonance suggesting that a respondent is under conflict when making a choice. When under conflict, respondents are more likely to select the no-choice option that takes shares from the other alternatives. “Free choice (having the no-choice option) provides an alternative route for reducing the psychological discomfort associated with forced choice under preference uncertainty. If the option not to choose is unavailable consumers resolve the forced choice problem by selecting other options that are associated with the least potential for significant error” (Dhar & Simonson 2003, p. 148).

Dhar and Simonson (2003) researched the difference in the no-choice option between a single-stage free choice task, where “none” is another alternative (traditional “none”), and a two-stage task, where a forced choice is followed by no-choice option (dual response). Respondents in the single-stage condition were presented with two or three product alternatives and also presented, as another alternative, the option to defer their product choice and go to another store. Conversely, respondents in the two-stage condition were first asked to choose between two or three products, without the no-choice option, hence a forced choice. After making the forced selection they were asked if they would remain with their choice or prefer to go to another store, hence, deferring their option. Across all the tested product categories 31% more respondents chose to defer choice in the two-stage condition than in the single-stage condition. For example, in the microwave oven category 10% of the respondents deferred choice in the single stage free choice task compared to 35% who selected to defer choice in the two-stage task. Although this trend is consistent across all product categories, the magnitude difference in choice deferral between the two conditions varies, most likely as a function of product involvement.

To replicate the results of Dhar and Simonson (2003), the second hypothesis is proposed:

**H2**: Dual-response (a forced choice is followed by a no-choice option) will have higher proportion of choice deferral than the traditional “none” (single-stage free choice task where “none” is another alternative).

When comparing the traditional “none” to the dual-response task, Dhar and Simonson (2003) explained the greater degree of choice deferral in the dual-response task as the respondents’ strategy of alleviating the psychological discomfort caused by a forced choice. They propose that this assertion might appear inconsistent with the endowment effect where selected options are overvalued because this effect does not apply to psychological discomfort generated by forced choice. Although this is a plausible explanation when compared to the single-stage free choice task, cognitive dissonance theory would predict that respondents would choose to overvalue their selection as a means of reducing the discrepancy between a behavior and subsequent evaluation, hence, reduce choice deferral. The assertion by Dhar and Simonson (2003) that there is no reduction of cognitive dissonance in a forced choice task followed by evaluation suggests the need further investigation.

Given respondents’ strategy of choosing the no-choice option as a means of alleviating discomfort when not fully committed to a selection in a forced choice task suggests that this effect of discomfort should dissipate with time, hence the respondent being more likely to support the forced selection after a time delay. Dhar and Simonson (2003) assigned respondents to either a single-stage or two-stage stage decision task. Those in the two-stage decision task with time delay chose among the items in the forced choice task, then participated in a filler task that took five minutes and then returned to the original task, evaluating the items chosen to decide if to remain with these items or go to another store to purchase another product. As expected, choice deferral was greater in the dual-stage task, when respondents were given the no choice option immediately after being forced to choose, than in the single stage decision where the “none” option was another alternative. However, the results show that time delay in the dual-response task decreased the level of deferral to the degree that the share of the “none” alternative was not significantly different than the single stage task. A plausible explanation is that given enough time to distance the selection from the context of the competitive set, respondents no longer experience discomfort from a forced choice. Hence, the two-stage forced choice task with time delay is analogous to the single-stage task with free choice in terms of choice deferral.

So far the single-stage free choice, where “none” is another alternative (traditional “none”) has been compared to the two-stage forced choice, where forced selection of an alternative is followed by a willingness to purchase question (dual response). According to Dhar and Nowlis (2004), in the real world, the consumer might be faced with a buy/non-buy decision followed by a selection decision regarding which option to buy.

> *Specifically, when the buy/no-buy decision is the initial focus (referred to as buy/no-buy response mode), consumer decision processes are more likely to be characterized by alternative-based evaluations (whether an option is acceptable). In contrast, consumers who are in an unconditional brand-choice response mode (the consumer has the option not to choose) are more likely to compare rival brands with each other, a process that results in more attribute based evaluations. (p. 423)*

It is our interpretation that the sequence of evaluation and selection within this two-phase process could be context dependent or due to the level of category involvement, each yielding a different cognitive strategy with different results regarding choice deferral as well as the weight assigned to the features of the product alternatives.

To illustrate that different decision processes yield different outcomes of choice deferral, Dhar and Nowlis (2004) offered two items, each option had three unique good features and three shared bad features. These items were offered across three between-subject conditions: The buy/no-buy condition presented the willingness to purchase option (choice deferral) before the selection. The unconditional brand choice option offered the option not to choose as another alternative (traditional “none”). The modified buy/no-buy condition is a delayed dual response in which respondents were forced to choose between two items in each of the three categories and after a time delay they evaluated their selections in terms of keeping the item or going to another store or location. Averaged across three categories (restaurants, vacations, apartments), 53% of respondents deferred choice in the buy/no-buy condition, 37% of respondents deferred choice in the traditional “none” condition, and 32% of respondents deferred choice in the modified buy/no-buy condition (delayed dual response). As in the previous study by Dhar and Simonson (2003), the delayed dual response did not differ from the traditional “none.” Of particular interest is that the respondents in the buy/no-buy condition had the greatest degree of choice deferral. Dhar and Nowlis (2004) proposed that the buy/no-buy condition, where evaluation precedes selection, facilitates the use of a category reference to evaluate the options. Conversely, in the traditional “none” condition the focus is on the differences among the options and the category reference information is less relevant on selection.

Dhar and Nowlis (2004) further tested the category reference effect in both the buy/no-buy condition and single stage with none option (traditional “none” condition) by providing respondents with either a superior or inferior category reference scenario before conducting the choice task. The results of the first study indicate that when provided with a superior category reference (when the options are inferior to the reference category), 39% of respondents deferred choice in the traditional "none" condition while 62% of respondents deferred choice in the buy/no-buy condition. Furthermore, when the category reference is inferior to the alternatives respondents in the buy/no-buy condition are less likely to defer choice than respondents in the traditional "none" condition. This finding supports that in the buy/no-buy condition, when evaluation precedes selection, the respondents are using category reference as a baseline for choice deferral.

Dhar and Nowlis (2004) show that choice outcomes are affected by the order of the buy/no-buy decision and selection. Interestingly, their study did not compare the buy/no-buy condition to the dual-response condition in which the only difference is the sequence of the forced selection and evaluation (choice deferral question). It has been shown that the respondent utilizes category reference when the buy/no-buy decision precedes selection, however, an opportunity of additional research is to further examine the implication of a traditional dual-response task where selection precedes the choice deferral question. Before expanding upon the research of Dhar and Nowlis (2004), it was first necessary to replicate their findings, however, the relationship of the reference category to the scenarios was not controlled such that an overt category reference was not offered.

Based on the strategy of category reference in the buy/no-buy dual task condition (we refer to this condition as the reversed dual response), we propose our third hypothesis:

**H3**: Reversed dual response, where evaluation precedes selection, will have higher proportion of choice deferral than the traditional "none" (single-stage free choice task).

Drawing upon the research of Leon Festinger (Festinger 1955, as cited in Brehm 1956) on "the relation between cognition and action," Brehm (1956) further examined "post decision changes in the desirability of alternatives" as it applied to consumer research. Specifically, Brehm (1956) hypothesized the following:

> *1. Choosing between two alternatives creates dissonance and a consequent pressure to reduce it. The dissonance is reduced by making the chosen alternative more desirable and the unchosen alternative less desirable after the choice than they were before it. 2. The magnitude of the dissonance and the consequent pressure to reduce it are greater the more closely the alternatives approach equal desirability. (p. 384)*

In the context of consumer behavior, Brehm (1956) had respondents rate 8 products (an automatic coffee-maker, an electric sandwich grill, a silk-screen reproduction, an automatic toaster, a fluorescent desk lamp, a book of art reproductions, a stop watch, and a portable radio) in terms of usefulness (desirability). To manipulate the degree of dissonance, subjects were presented with two products from which they can keep one. In the high dissonance condition the second product was as highly desirable as the first, in the medium dissonance condition the second product was moderately lower in desirability than the first, and in the low dissonance condition the second product was much less desirable than the first. After making a choice between the two products, respondents were asked to rate the products again, hence, allowing to measure the change of desirability for the chosen and unchosen products as well as the products not involved in the choice. Change in dissonance was measured by comparing the change between the first and second rating of the chosen and unchosen products. As expected, there was a reduction of cognitive dissonance by increasing the desirability of the chosen product and decreasing the desirability of the unchosen product. Furthermore, the degree of dissonance reduction was greater when both products were equal in desirability.

It has always been the case that to reduce discomfort within the individual there needs to be agreement between thoughts, emotions, and actions. An advancement since Festiner's (1957) original work is that dissonance reduction has been found to be multi-faceted and rather complex throughout the decision process. Sweeney, Hausknecht, and Soutar (2000) have improved the traditional desirability measurement by identifying three dimensions of dissonance, one emotional and two cognitive (wisdom of purchase and concern over deal) in the development of a multidimensional scale to measure cognitive dissonance after purchase. Additionally, Sweeney *et al.* (2000) cited purchase criteria that must exist for dissonance to occur with an important criterion being "the consumer must feel free in making the choice. That is, the decision must be made voluntarily" (p. 374). It could be argued that a forced choice task, not giving the respondent the "none" option, is indeed a free choice because the consumer chooses freely among the provided options. That is, the respondent is engaging in cognitive process to evaluate the merit of the selection such that there is personal responsibility towards justifying the choice. In addition, it is suggested that in a real-world situation, where the consumer has the option to defer choice, the choice is still forced, hindered by financial, time, and availability constraints.

Acknowledging previous work concerning the role of cognitive dissonance in consumer decision making towards high involvement products, where the person has more investment in the decision, Nordvall (2014) focused on low involvement purchase such as grocery shopping for organic food. Theoretically, regardless of product involvement, the same underlying mechanism should apply where attitudes and behaviors should be in agreement to avoid a state of stress. Participants were presented with equal numbers of organic and non-organic items on a computer screen. Each item was rated on a Likert scale from "never buy" to "buy sometimes" to "buy very often." For each pair of items the rating of each item indicated which items participants purchased to the same extent, to be considered equally attractive. Cognitive dissonance would result when both items are equally attractive, and there would be disagreement between attitudes and behavior. The participants chose between pairs of products rated similarly attractive and preferable, with each pair having one organic and one non-organic similar product. Participants were told to choose a product into the shopping basket from each pair. After selection the participants rated each product again. Along with the product was a reminder to the respondent if the product was chosen or rejected. The results show significantly higher score changes for the non-organic item after it was chosen, hence supporting dissonance reduction by raising the desirability of the chosen item. However, there should also be lowering of the desirability of the rejected item. No tendencies for dissonance reduction was found for the organic item when choosing the non-organic option. Basically, the organic items were not diminished in rating when rejected in favor of the non-organic items. Although not all hypotheses were supported for this low involvement product, in agreement with Brehm (1956), this study shows an inclination of the consumer to indicate an increased preference towards the chosen product as compared to preselection.

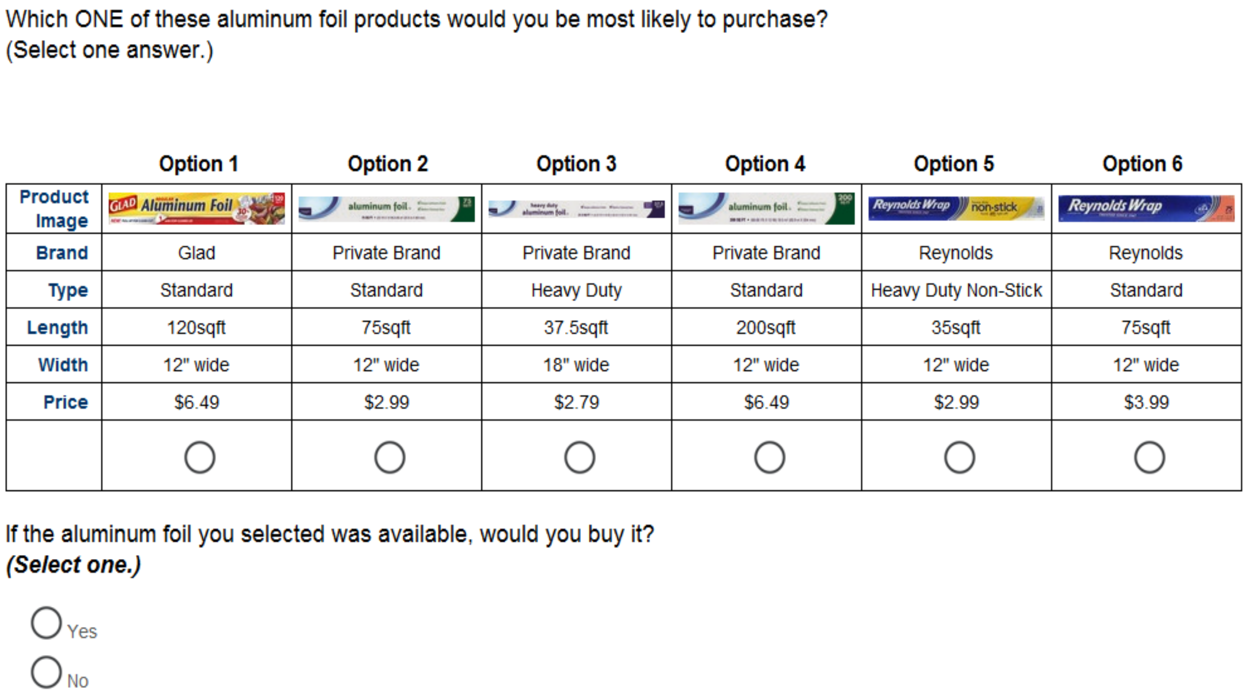
Research showing increased product favorability, hence the tendency to reduce choice deferral, after a selection has been made extends towards the traditional dual-response methodology in discrete choice analysis where selection precedes evaluation of willingness to purchase. It is suggested that traditional dual response creates a situation of cognitive dissonance within the individual that is diminished by reducing choice deferral, hence increasing willingness to purchase the product just selected. Drawing upon reduction of cognitive dissonance theory (Brehm 1956, Festinger & Carlsmith 1959) as a mechanism of reducing stress by changing thoughts to be consistent with behaviors, specifically as it applies to consumer behavior, we propose the fourth hypothesis:

**H4**: Reversed dual response, where the opportunity to defer choice precedes selection, will have higher proportion of choice deferral than the traditional dual response, where selection precedes the opportunity to defer choice.

## Participants and Procedure

A nine-minute online study was conducted with 1,201 respondents. Respondents were primary grocery shoppers age 24 to 65, who buy aluminum foil at least once every six months. Two attributes were tested, product and price, with each product having five unique price points (alternative-specific design). We used discrete choice methodology with different versions of the "none" alternative (four conditions with each condition having approximately 300 respondents). Respondents rated 13 scenarios, with one of the scenarios used as a holdout to test the accuracy of the model. Figures 1 through 4 illustrate the respondent task in each condition.

**Figure 1. *Traditional Dual Response* with forced choice followed by willingness to purchase question.**



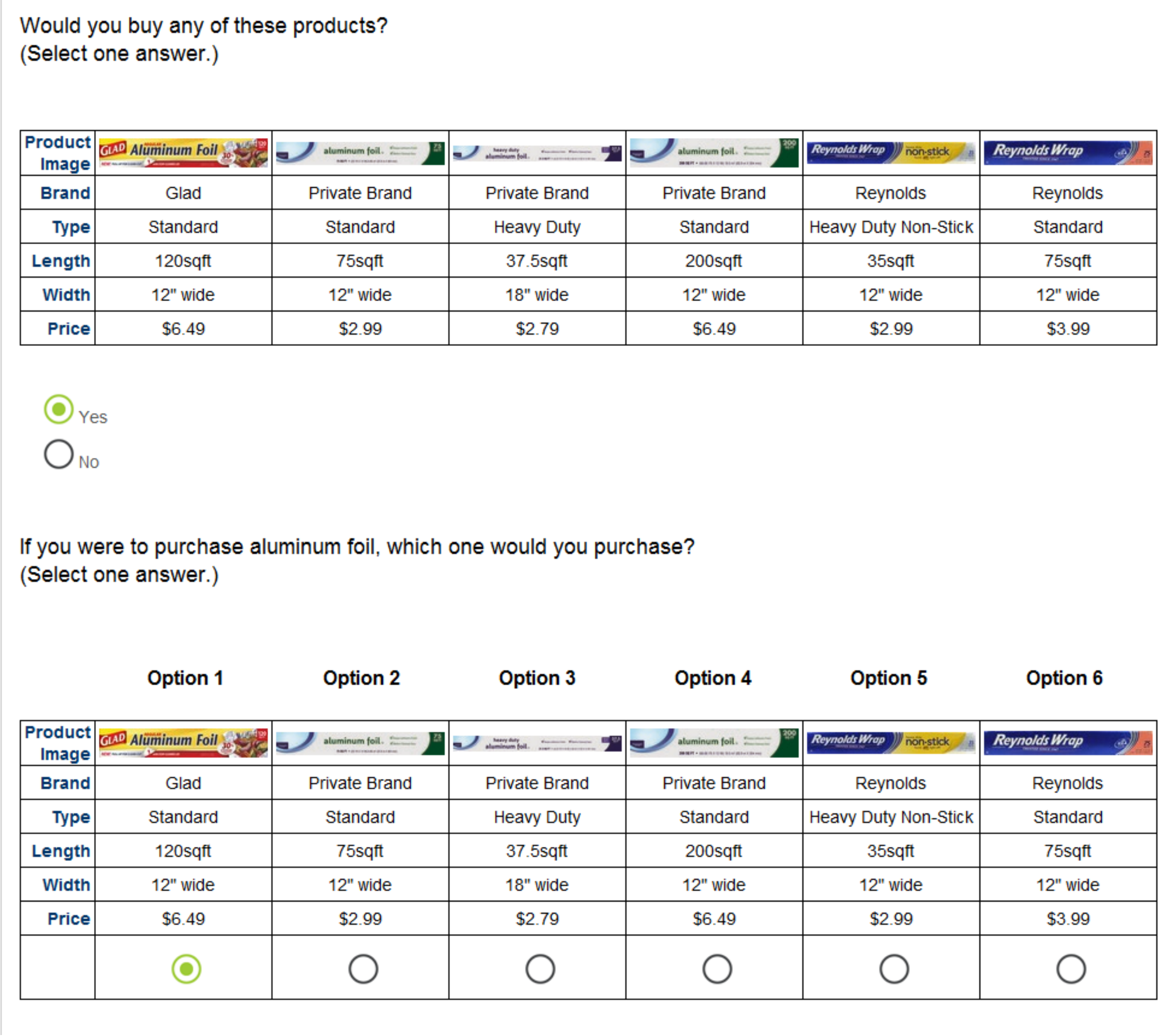
Which ONE of these aluminum foil products would you be most likely to purchase?
(Select one answer.)

| | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 | Option 6 |
|---|---|---|---|---|---|---|
| **Product Image** | | | | | | |
| **Brand** | Glad | Private Brand | Private Brand | Private Brand | Reynolds | Reynolds |
| **Type** | Standard | Standard | Heavy Duty | Standard | Heavy Duty Non-Stick | Standard |
| **Length** | 120sqft | 75sqft | 37.5sqft | 200sqft | 35sqft | 75sqft |
| **Width** | 12" wide | 12" wide | 18" wide | 12" wide | 12" wide | 12" wide |
| **Price** | $6.49 | $2.99 | $2.79 | $6.49 | $2.99 | $3.99 |
| | ○ | ○ | ○ | ○ | ○ | ○ |

If the aluminum foil you selected was available, would you buy it?
***(Select one.)***

○ Yes
○ No

**Figure 2.** ***Reversed Dual Response*** **with Buy/No-Buy question followed by forced choice. The second grid appears on the same screen** ***after*** **the first question is answered. Regardless if the answer is yes or no, the follow up grid will** ***always*** **appear.**



Would you buy any of these products?
(Select one answer.)

| Product Image | | | | | | |
|---|---|---|---|---|---|---|
| Brand | Glad | Private Brand | Private Brand | Private Brand | Reynolds | Reynolds |
| Type | Standard | Standard | Heavy Duty | Standard | Heavy Duty Non-Stick | Standard |
| Length | 120sqft | 75sqft | 37.5sqft | 200sqft | 35sqft | 75sqft |
| Width | 12" wide | 12" wide | 18" wide | 12" wide | 12" wide | 12" wide |
| Price | $6.49 | $2.99 | $2.79 | $6.49 | $2.99 | $3.99 |

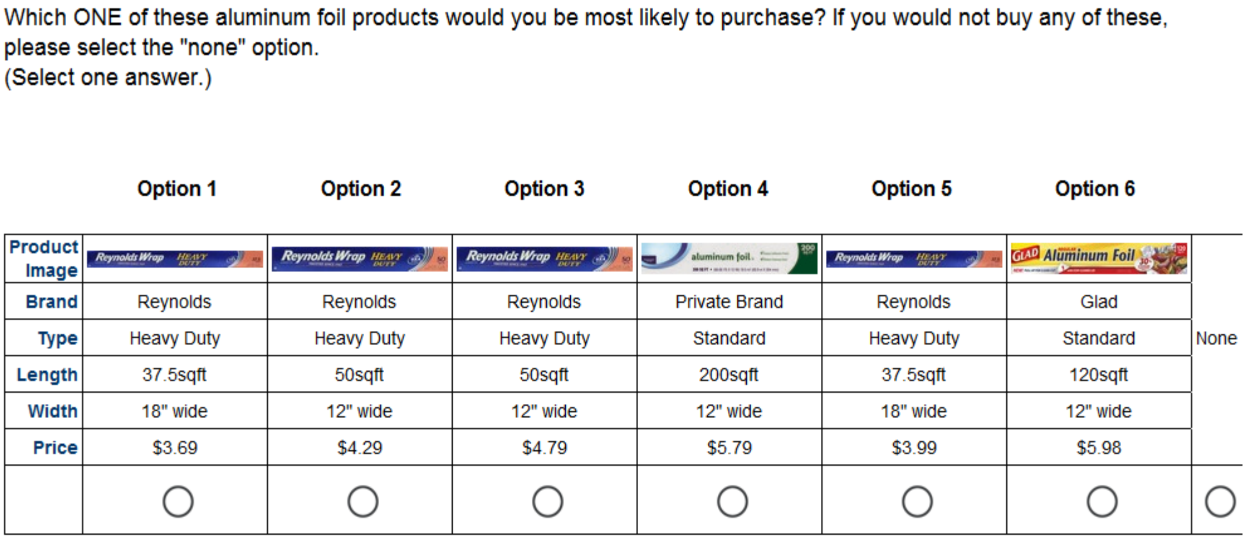
● Yes

○ No

If you were to purchase aluminum foil, which one would you purchase?
(Select one answer.)

| | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 | Option 6 |
|---|---|---|---|---|---|---|
| Product Image | | | | | | |
| Brand | Glad | Private Brand | Private Brand | Private Brand | Reynolds | Reynolds |
| Type | Standard | Standard | Heavy Duty | Standard | Heavy Duty Non-Stick | Standard |
| Length | 120sqft | 75sqft | 37.5sqft | 200sqft | 35sqft | 75sqft |
| Width | 12" wide | 12" wide | 18" wide | 12" wide | 12" wide | 12" wide |
| Price | $6.49 | $2.99 | $2.79 | $6.49 | $2.99 | $3.99 |
| | ● | ○ | ○ | ○ | ○ | ○ |

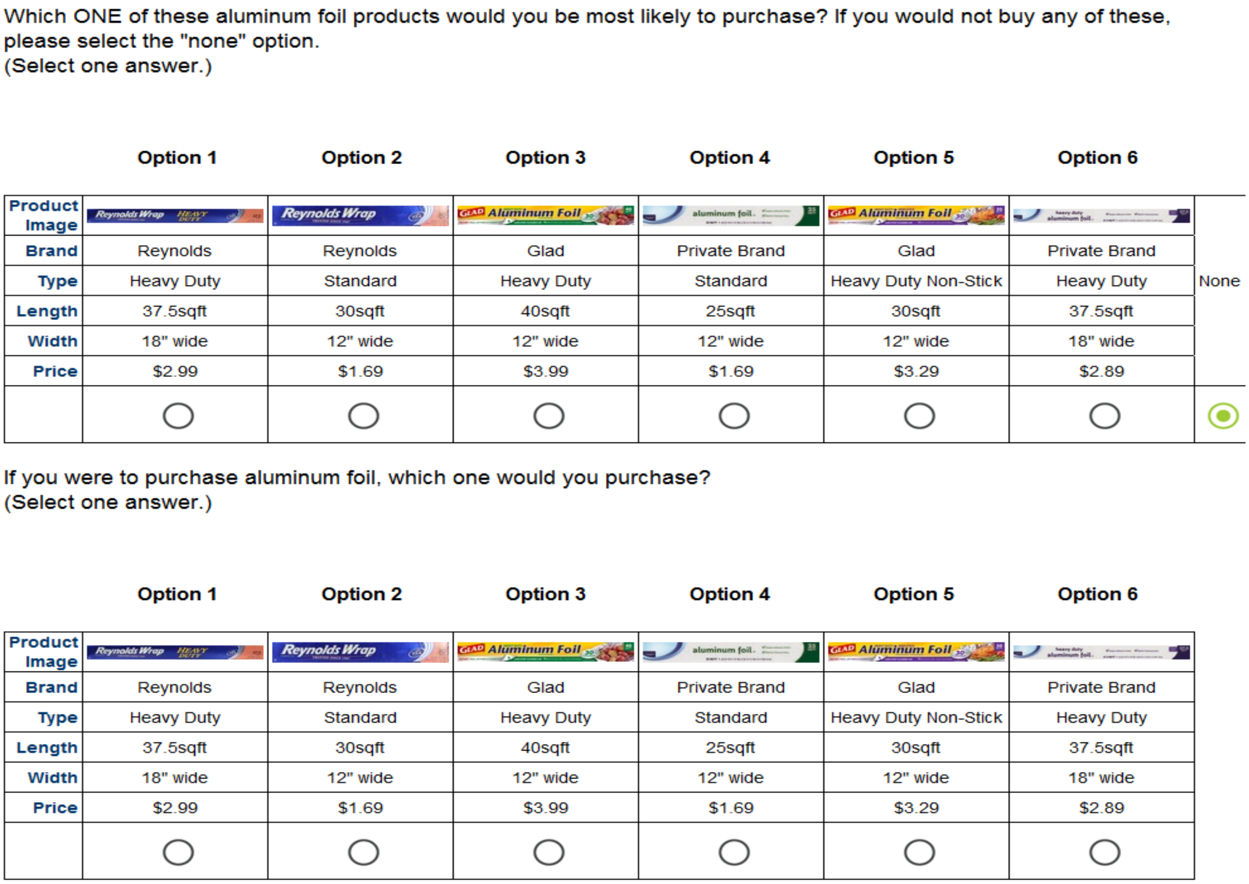
**Figure 3.** ***Traditional "None"*** **where the "None" alternative is another option.**



Which ONE of these aluminum foil products would you be most likely to purchase? If you would not buy any of these, please select the "none" option.
(Select one answer.)

| | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 | Option 6 | |
|---|---|---|---|---|---|---|---|
| Product Image | | | | | | | |
| Brand | Reynolds | Reynolds | Reynolds | Private Brand | Reynolds | Glad | |
| Type | Heavy Duty | Heavy Duty | Heavy Duty | Standard | Heavy Duty | Standard | None |
| Length | 37.5sqft | 50sqft | 50sqft | 200sqft | 37.5sqft | 120sqft | |
| Width | 18" wide | 12" wide | 12" wide | 12" wide | 18" wide | 12" wide | |
| Price | $3.69 | $4.29 | $4.79 | $5.79 | $3.99 | $5.98 | |
| | ○ | ○ | ○ | ○ | ○ | ○ | ○ |

**Figure 4.** ***Traditional "None" with follow up*** **product selection when "None" is chosen. The second grid appears on the same screen only when the "None" alternative is selected. If "None" is not selected, the follow up grid is skipped.**



Which ONE of these aluminum foil products would you be most likely to purchase? If you would not buy any of these, please select the "none" option.
(Select one answer.)

| | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 | Option 6 | |
|---|---|---|---|---|---|---|---|
| Product Image | | | | | | | |
| Brand | Reynolds | Reynolds | Glad | Private Brand | Glad | Private Brand | |
| Type | Heavy Duty | Standard | Heavy Duty | Standard | Heavy Duty Non-Stick | Heavy Duty | None |
| Length | 37.5sqft | 30sqft | 40sqft | 25sqft | 30sqft | 37.5sqft | |
| Width | 18" wide | 12" wide | 12" wide | 12" wide | 12" wide | 18" wide | |
| Price | $2.99 | $1.69 | $3.99 | $1.69 | $3.29 | $2.89 | |
| | ○ | ○ | ○ | ○ | ○ | ○ | ● |

If you were to purchase aluminum foil, which one would you purchase?
(Select one answer.)

| | Option 1 | Option 2 | Option 3 | Option 4 | Option 5 | Option 6 |
|---|---|---|---|---|---|---|
| Product Image | | | | | | |
| Brand | Reynolds | Reynolds | Glad | Private Brand | Glad | Private Brand |
| Type | Heavy Duty | Standard | Heavy Duty | Standard | Heavy Duty Non-Stick | Heavy Duty |
| Length | 37.5sqft | 30sqft | 40sqft | 25sqft | 30sqft | 37.5sqft |
| Width | 18" wide | 12" wide | 12" wide | 12" wide | 12" wide | 18" wide |
| Price | $2.99 | $1.69 | $3.99 | $1.69 | $3.29 | $2.89 |
| | ○ | ○ | ○ | ○ | ○ | ○ |

## Results and Discussion

To evaluate the effect of how the "none" alternative is asked in a discrete choice task on choice deferral, for each respondent the mean aggregated "none" response across the twelve scenarios was entered into a one-way analysis of variance. An alpha level of .05 was used for all statistical tests. There was a significant main effect for choice deferral condition, F (3, 1197) = 19.47, p = .001.

Consistent with previous research, Hypothesis 1 has been supported (Table 1). Further planned pairwise comparisons using the Bonferroni procedure revealed a significant difference (p = .02) between choice deferral in the traditional "none" with follow-up (Mean (M) = 3.9%, Standard Error (SE) = .0138) compared to choice deferral in the traditional "none" (M = 7.9%, SE = .0138). When choice deferral was followed by a forced choice, respondents adopted the strategy of choosing a viable product thereby treating the free choice task as a forced choice and avoiding the follow-up question. Given that choosing the "none" alternative is a choice deferral strategy to reduce discomfort in the presence of a difficult decision between alternatives (Tversky and Shafir 1992), when choice deferral did not show to be a viable way of resolving this difficult decision due to a follow-up forced choice question, the respondents abandoned the choice deferral strategy in the free choice task and treated the task as a forced choice.

Inconsistent with previous research, Hypothesis 2 has not been significantly supported but the results are in the correct direction (Table 1). Further planned pairwise comparisons using the Bonferroni procedure revealed that choice deferral in the traditional dual response (M = 9.6%, SE = .0138) was not significantly different than choice deferral in the traditional "none" (M = 7.9%, SE = .0138). Dhar and Simonson (2003) believe that in the free choice task (traditional "none"), consumers experience a greater degree of commitment to an option when viewed as chosen freely in the presence of a no choice option. Conversely, when forced to make a choice and then given the opportunity to defer that choice, consumers are more likely to take advantage of this offer when compared to the free choice task. In could be that not being able to replicate the results of Dhar and Simonson (2003) is due to the use of a low involvement product in the current study that results in a smaller effect size for choice deferral. The result being in the correct direction with a low involvement product is promising. In a high involvement category, where choice deferral is assumed to be more pronounced, it would be easier to identify significant differences. It is important to note that a greater degree of choice deferral in the dual-response task than the single stage free choice task, as a way to reduce psychological discomfort generated by forced choice, is not contradictory to reduction of cognitive dissonance theory because free choice can be viewed as a more efficient path towards reduction of cognitive dissonance, as evidenced by greater willingness to purchase when the choice is made freely, in the presence of a "none" alternative.

Consistent with previous research, Hypothesis 3 has been supported (Table 1). Further planned pairwise comparisons using the Bonferroni procedure revealed a significant difference (p = .001) between choice deferral in the reversed dual response, where the buy/no-buy decision is made before forced choice, (M = 14.3%, SE = .0138) compared to choice deferral in the traditional "none" condition (M = 7.9%, SE = .0138). Dhar and Nowlis (2004) explain this finding in terms of alternative based evaluations relative to a category reference when the buy/no-buy decision is made prior to selection. More specifically, when the category reference is better than the choice alternatives, choice deferral is greater in the buy/no-buy condition than in the traditional “none” condition. Conversely, when the category reference is worse than the alternatives, choice deferral in the buy/no-buy condition is lower than in the traditional “none” condition. It should be noted that in the current study of a low involvement product a category reference was not provided to the respondents as a primer to the reversed dual response. Thus, the higher proportion of deferred choice in the reversed dual response compared to the traditional “none” condition could mean that, on average, respondents referenced a better category than the alternatives or another cognitive process can be at play in the absence of a predefined reference category.

Hypothesis 4 has been supported (Table 1). Further planned pairwise comparisons using the Bonferroni procedure revealed a significant difference (p = .004) between choice deferral in the reversed dual response (M = 14.3%, SE = .0138) and choice deferral in the traditional dual response (M = 9.6%, SE = .0138). The only variable manipulated between the reversed dual response and the traditional dual response was the order of selection and choice evaluation. Given the findings of a lower degree of choice deferral in the traditional dual response than in the reversed dual response, it is suggested that a possible explanation is the reduction of cognitive dissonance in the traditional dual response condition where selection is made prior to choice evaluation (Brehm 1956, Festinger & Carlsmith 1959). In a situation of forced choice, when a selection is made, a respondent is more likely to rationalize the selection with an increased level of willingness to purchase, hence, a lower level of choice deferral. Such a dissonance between selection and evaluation is suggested not to exist when evaluation precedes selection, hence, a higher level of choice deferral.

**Table 1. Means of Means of Deferred Choice in Each Experimental Condition.**

| | Traditional Dual Response | Reversed Dual Response | Traditional “None” | Traditional “None” with follow-up |
|---|---|---|---|---|
| Proportion of Deferred Choice | 9.6% (.179) | 14.3% (.216) | 7.9% (.169) | 3.9% (.087) |

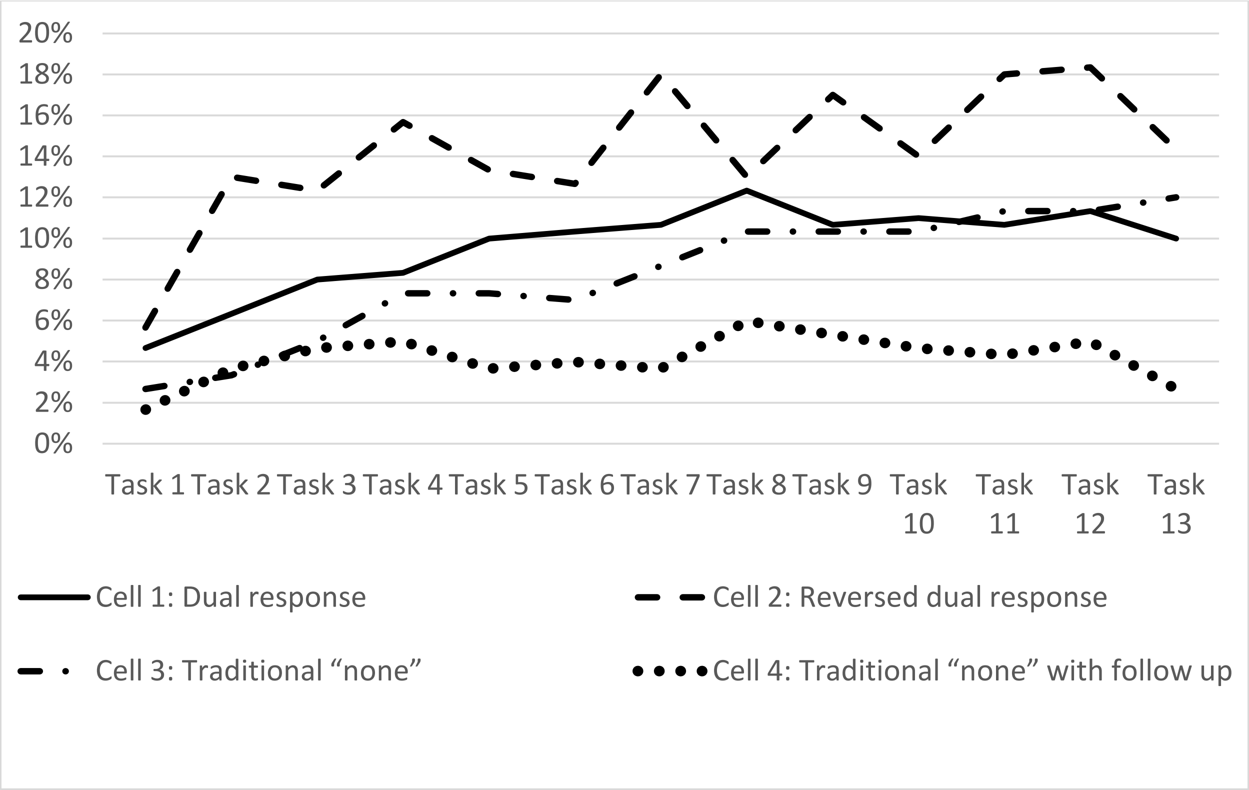
Note: Standard deviations are in parentheses.

Table 2 shows that for traditional dual response, reversed dual response, and traditional “none” there is an increase in choice deferral as the exercise progresses. This pattern of increased choice deferral as a function of scenario order is not as evident for traditional “none” with follow-up product selection. The explanation regarding a somewhat stable proportion of choice deferral in the traditional “none” with follow-up condition is that respondents are likely to have adopted a strategy of not deferring choice because when choosing the “none” alternative there was always a follow-up forced choice. The pattern of a steady and lower proportion of choice deferral across scenario order for traditional “none” with follow-up compared to the other three experimental conditions is further accentuated in Figure 5.

**Table 2. Proportion of Deferred Choice among the First and Last Three Scenarios between Conditions.**

| | Traditional Dual Response | Reversed Dual Response | Traditional "None" | Traditional "None" with follow-up |
|---|---|---|---|---|
| Proportion of Deferred Choice Earlier Tasks (first 3) | 6.3% | 10.3% | 3.7% | 3. 3% |
| Proportion of Deferred Choice Later Tasks (last 3) | 10.7% | 16.9% | 11.6% | 4.0% |

**Figure 5. Proportion of Choice Deferral as a Function of Scenario Order between Experimental Conditions.**



For each experimental condition, the duration of time to complete the discrete choice task was measured (Table 3). Both dual-response tasks took longer to complete compared to the traditional "none" tasks. The completion times of the tasks are as expected since a two-stage task has an additional follow-up question. However, reversed dual response is more enjoyable, on par with traditional "none," when compared to traditional dual response and traditional "none" with follow-up. It is suggested that the two most enjoyable tasks caused the least degree of stress to the respondent. The reversed dual response, where evaluation precedes selection, did not create a state of cognitive dissonance in the respondent and the traditional "none" was a free choice task since the respondent could readily choose the "none" alternative.

As seen in Table 3, we also evaluated the predictive accuracy of the models using holdout tasks and comparing each model's hit rate and mean absolute error (MAE). Moreover, although reversed dual-response model is slightly less predictive when compared to traditional "none" with follow-up, it performs on par with dual response and traditional "none," which are the most widely used methods today. While these finding are just directional, clear patterns can be observed.

**Table 3. Additional Measures between the Four Experimental Conditions.**

| | Condition 1: Dual response | Condition 2: Reversed dual response | Condition 3: Traditional "none" | Condition 4: Traditional "none" with follow-up |
|---|---|---|---|---|
| Time to complete | 5.6 min | 5.9 min | 4.2 min | 4.6 min |
| Top box enjoyment (4 point scale) | 32% | 36% | 37% | 34% |
| Hit Rate | 52% | 55% | 54% | 58% |
| MAE | 5.4 | 5.2 | 6.3 | 4.5 |

Ideally, the utility estimates for attribute levels should be equal in different exercises regardless if using the traditional "none" or a dual-response "none" alternative. That would be true if creating a forced-choice from the traditional free-choice task by deleting the "none" option did not impact the relative probability of choosing the remaining alternatives, and the addition of the "none" alternative in the second stage of the dual-response task drew proportionately from each alternative. But there is some evidence, as mentioned earlier, that especially in two attribute studies (product and price) adding and deleting the "none" does not draw proportionately from the other alternatives of the choice task due to compromise alternatives. This means that in some choice situations deleting the "none" option systematically violates the independence of irrelevant alternatives (IIA property), as suggested by Dhar and Simonson (2003).

When estimating the utilities using a multinomial logit, there is the assumption of IIA, however as noted, this IIA assumption is violated during the actual tradeoff task, especially as a function of methodological details during the tradeoff task. Since the "none" option does not draw proportionally from all alternatives, choice deferral could not be isolated from the violation of the IIA assumption. This results in similar but not identical utilities when deleting the "none" option for different "none" alternatives. As such, in accordance with differing proportions of deferred choice between the four experimental conditions, it also warrants that the feature utilities of the products will differ as well. Due to the ordinal nature of product specific pricing, the price attributes were not included in the correlation of the utilities between the experimental conditions. As can be seen in Table 4, correlations among the 17 product utilities, with and without the "none" utility, between each of the experimental conditions suggests a fair degree of similarity, most likely between the most and least preferred products. However, a less than perfect correlation suggests that preference shares between simulated results will not always agree, depending on how choice deferral is phrased and, of course, sample differences as well.

**Table 4. Correlation among the Utilities between the Experimental Conditions.**

| | Condition 1 vs Condition 2 | Condition 1 vs Condition 3 | Condition 1 vs Condition 4 | Condition 2 vs Condition 3 | Condition 2 vs Condition 4 |
|---|---|---|---|---|---|
| With “none” | .81 | .86 | .90 | .84 | .74 |
| Without “none” | .83 | .86 | .92 | .90 | .81 |

## GENERAL DISCUSSION

Decision making regarding a product or service, either low or high involvement, is a two-stage process of evaluation and selection, with the sequence of these aspects varying depending upon the situation. In the real world the consumer has the option of selecting from the current alternatives or deferring choice at the risk of losing some or all of the available options. In an effort to mimic a real-world situation the tradeoff task has evolved from rating a particular scenario in terms of likelihood to purchase to an actual choice task where respondents have the option to defer choice if indecisive regarding the current offerings.

The deferral option in the discrete choice task can be another alternative, as in the single-stage free choice task, (referred to as the traditional “none”) or a separate stage in the dual-response task that involves selection and choice evaluation. The buy/no-buy response can either precede or follow the forced selection stage. Hence, the “none” response used to estimate the proportion of consumers unwilling to purchase can be a result of a single-stage free choice task, a traditional dual-response task, or a reversed dual-response task, with each exercise yielding a different estimation of the proportion of choice deferral.

It has been suggested that the traditional dual response, selection precedes evaluation, is prone to reduction of cognitive dissonance. In an effort to align thought and action, the respondent will be more willing to indicate product purchase after that product selection has been made than in a situation where the selection has not yet occurred. At this time there is no definite conclusion regarding the order of selection and evaluation, with each strategy occurring under different circumstances and resulting in a different degree of choice deferral (the “none” utility in a multinomial logit model).

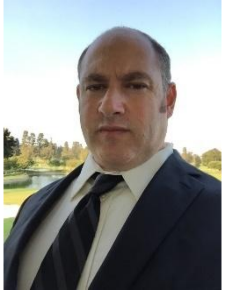
It has been found in this study and supported in the literature that not having identical utilities between the experimental conditions indicates that the “none” option impacts the utilities of the product or service feature levels disproportionately. This violation of independence of irrelevant alternatives, even in the absence of using the “none” alternative in simulations, suggests that choice deferral is not a trivial issue.

The “none” alternative is used by certain researchers to accurately size share of preference, or market share with proper calibration, of those who are not interested in the product line. The estimate of willingness to purchase has traditionally been overestimated in survey research, hence, requiring calibration during the modeling phase. Although calibration using external data sources can prove to be reliable in adjusting the stated preference estimates of both the feature levels and the “none” proportion, this adjustment method cannot always be implemented due to unavailability of external data especially in situations of new product development where such data do not exist. Hence, it is the goal of this research to develop a better estimation of choice deferral, especially in situations where external data do not exist.

## FUTURE RESEARCH

- Our “none” alterative exploration in this study was conducted in a low involvement product category. Further research is warranted in a high involvement product category, where choice deferral is higher. We assume, that the cognitive dissonance effect is much greater in high involvement categories.
- Additional research needs to be conducted with more than two attributes for the choice alternatives because with more attributes and levels there is usually no clear compromise. With less of a compromise effect in multi-attribute studies it should be possible to better isolate the choice deferral from the IIA violation than in studies with fewer attributes and levels.
- More research is required to examine the impact of the range effect on choice deferral. Schlereth and Skiera (2016) show that attribute ranges have a large impact on the no-purchase option in choice experiments. If, due to larger attribute ranges, multiple products are presented in a choice set then this increases the likelihood that an attractive product will be compared to the “none” option. Respondents with higher purchase probability in the product category will be more likely to choose the attractive product over the no-purchase option compared to respondents with lower purchase probabilities in the product category, whose preference of the attractive product is closer to the no-purchase option. Without taking the range effect into account, it is questionable which way of presenting the “none” option results in more accurate utilities.
- Dual response and reversed dual response will have higher proportion of deferred choice than the traditional “none” task. We should investigate this under the assumption that the “none” option does not draw proportionally from the other alternatives. This implies that we must address the IIA property by introducing a nested logit or mixtures of multivariate normal for the estimation of the part-worth utilities.
- Reversed dual response (where the buy/no-buy is offered before selection) has been shown to have a higher proportion of choice deferral than the traditional dual response where selection precedes the opportunity to defer choice. It is necessary to test the reversed dual-response assumption under different conditions of the market reference category. For example, examining the degree of choice deferral in a situation where a better reference category is provided, a worse reference category is provided, and where no reference category is provided.
- As the research industry is battling declining response rates, short attention spans, and struggling with attracting younger respondents, we need to continue creating smarter surveys. Surveys need to be short, feel intuitive, enjoyable, and free of discomforts such as cognitive dissonance. Surveys could be adaptive to avoid choice tasks dominated by one alternative. At the same time, we need to preserve predictive accuracy and produce purchase intent that is more in line with client expectations.

Ula Jones



Tomer J. Ozari

Peter Kurz

## REFERENCES

Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., & Rao, V. (1994). Combining revealed and stated preferences data. Marketing Letters, 5 (4), 335–350.

Brazell J. D., Diener C. G., Karniouchina E., Moore W. L., Séverin V., & Uldry P. F. (2006). The no-choice option and dual response choice designs. Market Lett, 17, 255–268.

Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. Journal of Abnormal Social Psychology,52, 384–389.

Breidert, C., Hahsler, M., & Reutterer, T. (2006). A review of methods for measuring willingness-to-pay. Innovative Marketing, 2 (4), 8–32.

Dhar, R., & Nowlis, S. M. (2004). To buy or not to buy: Response mode effects on consumer choice. Journal of Marketing Research, 41 (November), 423–432.

Dhar, R., & Simonson, I. (2003). The effect of forced choice on choice. Journal of Marketing Research, 40 (May), 146–160.

Festiner, L. (1957). A theory of cognitive dissonance. Stanford, CA: Stanford University Press.

Festinger, L. (1964). Conflict, decision and dissonance. Stanford, CA: Stanford University Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. Journal of Abnormal Social Psychology, 58, 203–210.

Green, P. E., & Rao, V. R. (1971). Conjoint measurement for quantifying judgmental data. Journal of Marketing Research, 8 (August), 355–363.

Karty, K. D., & Yu, B. (2012). Taking nothing seriously or “much ado about nothing.” Sawtooth Conference Proceedings Sequim, 129–151.

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), 350–367.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In: P. Zarembka (Ed.), Frontiers in econometrics (pp. 105–142). New York: Academic Press.

Nordvall, A. C. (2014). Consumer cognitive dissonance behavior in grocery shopping. International Journal of Psychology and Behavioral Sciences, 4 (4) 128–135.

Sawtooth Software (2013): The CBC System for Choice Based Conjoint Analysis Version 8. Sawtooth Software Inc. Orem Utah.

Schlereth, C., & Skiera, B. (2016). Two new features in discrete choice experiments to improve willingness to pay estimation that result in SDR and SADR: Separated (Adaptive) Dual Response. Management Science (in print).

Sweeney, J. C., Hausknecht, D., & Soutar, G. N. (2000). Cognitive dissonance after purchase: A multidimensional scale. Psychology & Marketing. 17 (5), 369–385.

Tversky, A., & Shafir, E. (1992). Choice under conflict: The dynamics of deferred decision. Psychological Science, 3 (6), 358–361.

# THE ART AND SCIENCE OF NESTED LOGIT: CASE STUDIES FROM MODELING MANY SKUS

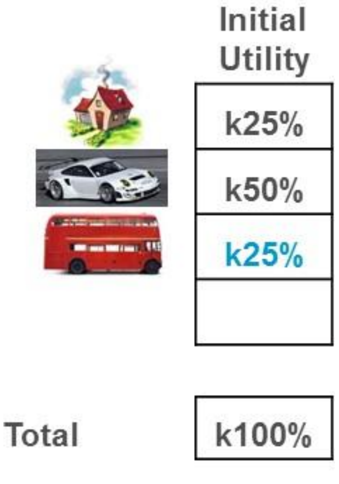
**_KEVIN LATTERY_**
_SKIM GROUP_

## 1.0 INTRODUCTION

Consumers often make choices by grouping similar items together. For example, consumers at a quick service restaurant may group the French Fries and other sides together. As one changes the price of a medium size French Fry, it is likely to impact whether one chooses other French Fries or maybe switches to Onion Rings. In contrast, changing a medium French Fry probably has little impact on the choice of coffee. In general the more similar items are perceived to each other, the more likely they impact each other when changed. In a conjoint study we call these similar products correlated alternatives.

The standard model in conjoint analysis is multinomial logistic regression (MNL). MNL assumes that there is no correlation among the alternatives. This is known as Independence of Irrelevant Alternatives (IIA). Obviously if we assume no correlation among alternatives when there is strong correlation, our models can be problematic.

This presence of correlated alternatives is a widely recognized problem. Perhaps the most famous example of this is the Red Bus/Blue Bus problem. The idea here is that the Red Bus and Blue Bus are perfectly correlated. Assume we initially have only 3 alternatives: home, car, or a red bus. The exponentiated utility of those three alternatives is proportional to their share.



When we introduce a blue bus, we assume it will have the same utility as the red bus—people most likely do not care about the bus being blue or red. They are perfectly correlated alternatives. Adding the blue bus alternative we get these results:

| | Initial Utility | Utility with Blue | New Share Assume IIA |
|---|---|---|---|
| | k25% | k25% | 20% |
| | k50% | k50% | 40% |
| | k25% | k25% | 20% |
| | | k25% | 20% |
| Total | k100% | k125% | 100% |

We now have 40% of people riding a bus (20% red and 20% blue). We could keep adding different color buses and further increase the percentage of bus riders. If we had 20 different colored buses, almost everyone would ride a bus.

In reality, we expect that the share for the blue bus would come solely from the red bus because they are correlated alternatives. In the above example this means red bus and blue bus should each be 12.5%.

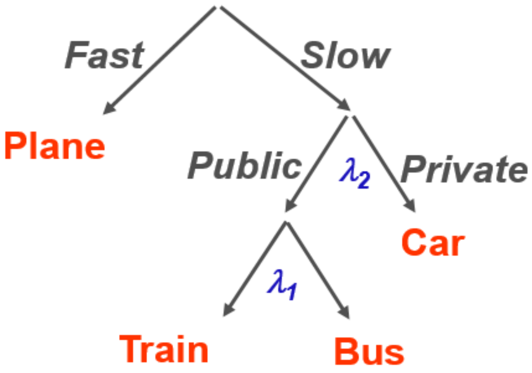
This kind of problem with correlated alternatives was especially prominent in the earlier days of conjoint when we developed aggregate models. With aggregate models the problem with correlated alternatives was clearly visible. Not accounting for correlation resulted in predictions that looked obviously wrong. In the early 2000's, respondent-level models such as Hierarchical Bayes (HB) became prominent. This made the problem far less visible. In a respondent-level HB model, IIA happened at the respondent level, but was significantly reduced when aggregating over respondents.

While respondent-level models reduce IIA in aggregation, in some cases the individual-level IIA trickles up to the aggregate level. We have seen this happen more frequently with conjoint studies showing shelf sets with many similar SKUs. Later we will show a detailed example. But the question remains as to what one should do when they are faced with sourcing that looks wrong. Some have suggested a duct tape post-hoc approach. That is we assume IIA during our estimation, but create predictions that take into account correlated alternatives. This means we estimate the model with one set of assumptions, and predict using another, which is clearly not ideal. Later we will see that this kind of mismatch between estimation and prediction lowers predictive validity.

In this paper we show how nested logit can be used during estimation and prediction to model the correlation between alternatives. The next section (2.0) will describe what nested logit is and how it models correlated alternatives. Section 3 will give hands-on advice about how to build and test the nesting structure, using a specific case study. In Section 4 we describe how to estimate nested logit models. We first describe estimation at the aggregate level, and then show how to extend this to Latent Class and HB.

## 2.0 Basics of Nested Logit

Nested logit involves adding a tree-like structure to the alternatives. A simple example of such a tree is the following:



One way to think about the tree structure above is as a sequence of decisions. The first decision is whether to take a plane (if traveling far) or slower (but more immediate) ground transport. Then within slow ground transport we consider public vs private. The diagram also shows two λ parameters. These parameters represent the degree of similarity between the items in the nest. So $\lambda_1$ represents the degree of similarity between train and bus.

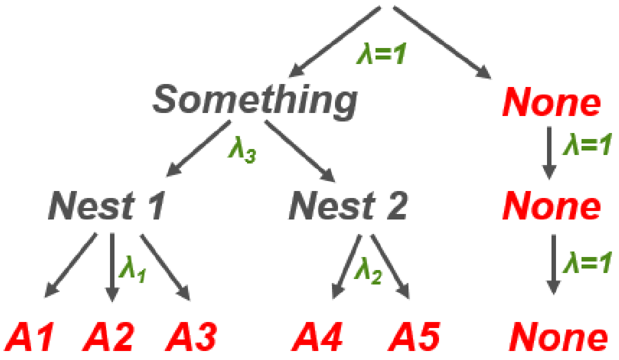
We typically define the λ parameters in the interval [0,1]. When λ =1, there is no correlation between the alternatives. If all the λ parameters in a nest structure are 1, then the nested structure is equivalent to the standard MNL. So mathematically, nested logit extends MNL, with additional λ parameters to model the correlation between alternatives grouped together in a nest or bundle. As λ moves from 1 toward 0, the alternatives are more similar to each other. As we approach 0, we get the red bus and blue bus which are perfectly correlated.

Section 2.1 is more technical, and will discuss the details of how nested logit computations are performed. Less technical readers may skip to section 3.0 without loss of understanding.

### 2.1 Mathematical Details of Nested Logit

There are a few variations of nested logit, but we will describe the most common version known as “Utility Maximization Nested Logit with Normalized Top Level.” The mathematics of nested logit works by estimating a utility for each nest and computing conditional probabilities. It is perhaps easiest to describe the calculations by focusing on a simple non-trivial example.

Assume we have the following nested structure:



At the bottom we have 5 alternatives A1 thru A5, and a None option. Estimation begins from the bottom-up. To illustrate the computations assume we know λ1 = .5 and λ2 = .25. Also assume we know the standard linear βx utility for each alternative. This is shown in Step 1 in the following table:

| | A1 | A2 | A3 | A4 | A5 | None |
|---|---|---|---|---|---|---|
| 1) V=βx | 0 | 0.5 | 1 | 0 | 0.7 | -2 |
| 2) V / λi | 0 | 1 | 2 | 0 | 2.8 | -2 |
| 3) e^Step2 | 1 | 2.7 | 7.4 | 1 | 16.4 | 0.14 |
| 4) Sum(Step 3) | 11.1 | | | 17.4 | | 0.14 |
| 5) λi * ln(step4) | 1.2 | | | 0.71 | | -2 |

Note: Values rounded to 1 decimal.

Step 2 simply divides each value in step 1 by its corresponding λ. So Nest 1 items are divided by .5 and Nest 2 items by .25. As λ goes to 0 this creates more extreme values, and for this reason we typically restrain λ to be at least 0.1. Step 3 exponentiates these utilities. These values replace the standard exponentiated utilities in MNL and are used for within-nest probabilities. Step 4 is also like MNL in that we sum the utilities, but only within each nest. So we have three separate sums corresponding to 3 base level nests.

Note that a nest with only one item (like None above) should have λ fixed at 1. Mathematically any value of λ will cancel itself out since step 3 and step 4 will be the same and give a probability of 1.

The ratio between step 3 and step 4 defines the probability within the nest. From the table above, the probability of A2 given Nest 1 is 2.7/11.1. So given Nest 1 we know the probabilities of each item in the nest. Likewise for Nest 2. We know the probabilities of each specific alternative A1–A5 *given the nest*. But what is the probability of Nest 1? Nest 2? None?

This is where Step 5 comes in. We take the natural log of step 4 and multiply that by its corresponding λ. This is called the inclusive value. It represents the (non-exponentiated) utility of the Nest. So the utility of Nest 1 = 1.2, Nest 2 = .71, and None = -2. These values then become Step 1 for the next level up. We repeat steps 2–5 again using these 3 values.

So given a set of base utilities V=βx for each alternative, a nesting structure, and λs for each nest we can compute utilities for each nest and corresponding probabilities. We then estimate the probabilities going down. So the probability of alternative A4 is the prob(Something) * prob(Nest2 | Something) * prob(A4 | Nest 2).

Obviously the calculations in a nested logit are much more involved than standard MNL. We have to estimate both the betas and λs. This also brings with it more difficulties in estimation. The preferred academic solution is to:

1. Estimate betas in a standard MNL.
2. Fix the betas and estimate the λs.
3. Taking 1 and 2 as starting points, estimate betas and λs simultaneously.

I find step 3 takes quite some time using optimization methods like BFGS or Newton-Raphson without specifying the gradient (which is difficult to do in a nested logit). So my preferred approach is to iteratively cycle through estimating betas and λs, conditional on previous results. After the initial betas in standard MNL, I use Powell's BOBYQA algorithm, as implemented in the R package minqa. It is fast and requires no gradients. I run this cycling between betas and λs until the log-likelihood converges. When running aggregate models I withhold sample from the estimation and base convergence on out-of-sample convergence.

## 3.0 How to Determine the Nesting Structure

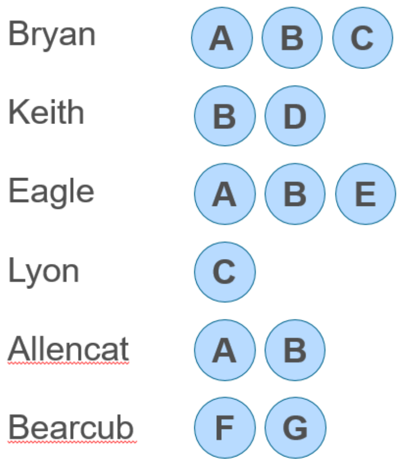
Nested logit requires a tree-like structure of the alternatives. But how does one develop this tree? Obviously one can specify any tree they like. But in this section we will suggest some specific data-driven methods to develop and test the tree.

To clarify our approach we introduce the following case study. The conjoint design was a shelf set. There were a total of 57 SKUs in the study, but each shelf showed 38 out of the 57 SKUs. Prices were varied and 1,157 respondents were allowed to choose multiple products. Each respondent evaluated 16 shelf sets.

The SKUs contained many products that varied by size or design. Some products were clearly close substitutes for each other, and it seems likely that there is indeed some degree of correlation among alternatives. A marketing expert in the specific category may have a good idea what products tend to be more similar to others. Here I show one way to develop the structure with empirical data.

### 3.1 Using Cross-Purchase Overlap to Develop Initial Structure

In most conjoint studies respondents evaluate several screens. In our case, each respondent saw 16 shelf sets. For each respondent we can then list all the alternatives they chose across all 16 tasks. For instance, let's assume across the conjoint tasks the respondents below chose the following items:



For any given pair of items, we can compute what we will call the overlap matrix. Consider the pair of items A, B. We first consider all the people who chose either A or B. This is 4 respondents, all but Lyon and Bearcub. Among those 4 respondents we then count the number who chose A & B. Three out of the 4 respondents did this, all but Keith. The ratio ¾ is the overlap percentage. It will always be between 0 and 1. More succinctly, the overlap is (Num of respondents who chose A&B)/ (Num of respondents who chose A or B).

For every given pair of items, we can compute the overlap, and create a symmetric matrix of overlap percentages for each item. In our case we have 57 SKUs and a 57 x 57 overlap matrix. We then convert this symmetric overlap to a symmetric matrix of distances. We want items with higher overlap to be closer together. We used Euclidean distance of (1-overlap). In some cases other transformations may be helpful to create more discrimination.

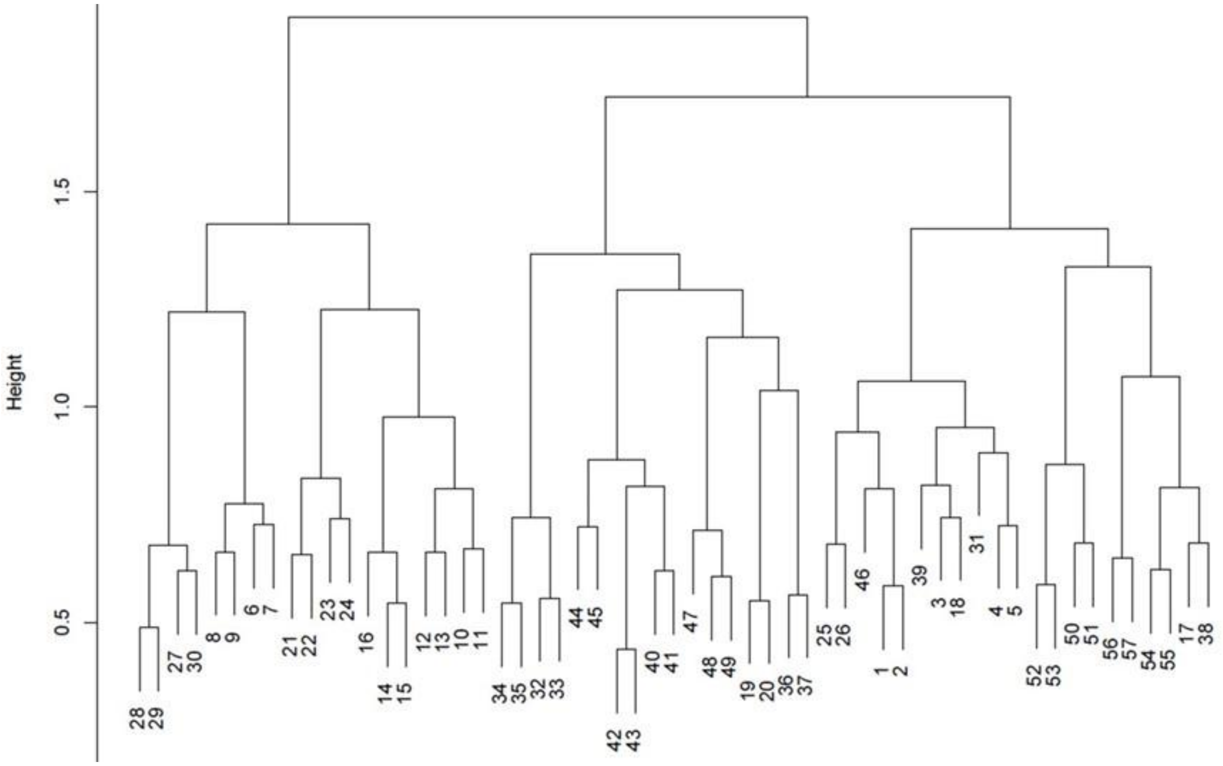
Once we have distances, we can do hierarchical clustering. In our experience Ward’s method works well. The R code looks like this, where overlap_p is the 57 x 57 overlap matrix:

```
kldist <- as.dist(1 - overlap_p)

klclust <- hclust(kldist, method = “ward.D2”)

plot(klclust)
```

This produces a dendrogram that can be exported to a PDF and looks like this:

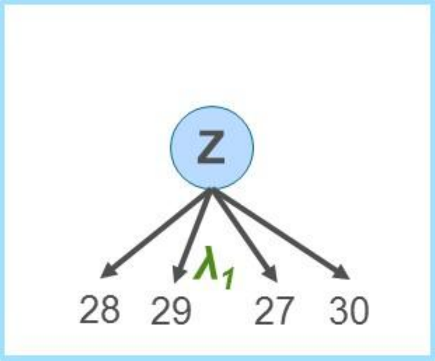


This dendrogram is meant to be a diagnostic tool showing which items are closest together (highest overlap). It gives you some direction about how the nesting structure might look. To further explore this we sort the overlap matrix in the same order as the dendrogram.

The first 8 SKUs are shown below. One can already see higher the overlap within these highlighted blocks. There is even stronger difference as we move further away. These two highlighted areas will be put into a nest higher up.

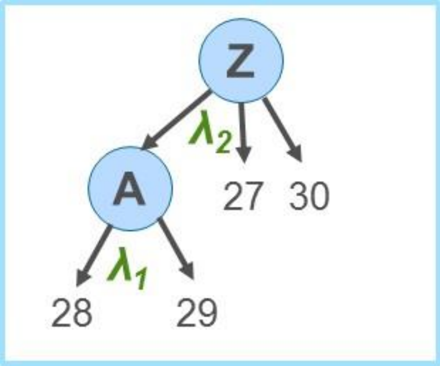
| Brand | Sub-Brand | Size | Style | Design | **28** | **29** | **27** | **30** | 8 | 9 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Main | Main w/New | 6 | R | A | **100.0%** | **51.2%** | **34.6%** | **39.6%** | 19.1% | 15.8% | 18.2% | 16.7% |
| Main | Main w/New | 6 | L | A | **51.2%** | **100.0%** | **36.6%** | **41.0%** | 17.7% | 15.4% | 16.5% | 17.3% |
| Main | Main w/New | 3 | G | A | **34.6%** | **36.6%** | **100.0%** | **38.1%** | 12.8% | 13.4% | 14.7% | 12.9% |
| Main | Main w/New | 6 | L | A | **39.6%** | **41.0%** | **38.1%** | **100.0%** | 17.0% | 16.6% | 17.2% | 16.1% |
| Main | Primary | 6 | SM | A | 19.1% | 17.7% | 12.8% | 17.0% | **100.0%** | **33.7%** | **25.2%** | **25.7%** |
| Main | Primary | 6 | B | A | 15.8% | 15.4% | 13.4% | 16.6% | **33.7%** | **100.0%** | **23.5%** | **31.3%** |
| Main | Primary | 6 | L | A | 18.2% | 16.5% | 14.7% | 17.2% | **25.2%** | **23.5%** | **100.0%** | **27.4%** |
| Main | Primary | 6 | S | A | 16.7% | 17.3% | 12.9% | 16.1% | **25.7%** | **31.3%** | **27.4%** | **100.0%** |

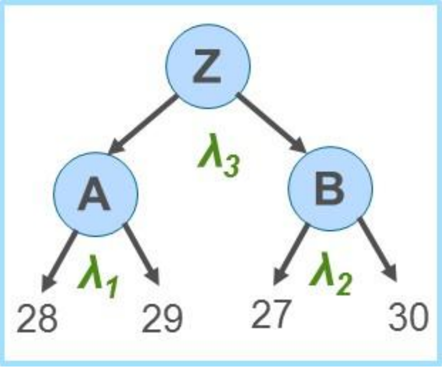
Viewing the overlap matrix sorted allows us to make decisions about which items belong in the same nest. If we take just the top 4 x 4 from above, we could put all 4 items together in one nest immediately like this:



| Brand | Sub-Brand | Size | Style | Design | 28 | 29 | 27 | 30 |
|---|---|---|---|---|---|---|---|---|
| Main | Main w/New | 6 | R | A | **100.0%** | **51.2%** | **34.6%** | **39.6%** |
| Main | Main w/New | 6 | L | A | **51.2%** | **100.0%** | **36.6%** | **41.0%** |
| Main | Main w/New | 3 | G | A | **34.6%** | **36.6%** | **100.0%** | **38.1%** |
| Main | Main w/New | 6 | L | A | **39.6%** | **41.0%** | **38.1%** | **100.0%** |

Or we could build the nest up with the first two and last two. Below are two examples of the nesting structures that could apply:





So how are we to know which of these nesting structures to use? Of course one can make decisions based upon knowledge of the category. And indeed we would rarely build a nesting structure that is contrary to expert knowledge. But we do not have to rely upon expert judgment. We can actually test the different models to determine which is best.

### 3.2 Testing Different Nesting Structures

For the sake of simplicity we usually test these different models at the aggregate level. This is similar to what is typically done in menu-based conjoint. With menu-based conjoint, we take each specific item and specify which other items impact it. For example we might specify that French fries are impacted by other fries and sides, but not coffee. This is similar to specifying nests, and in fact some people have suggested that one should do menu-based conjoint with nesting. In any case, one typically tests menu specification at the aggregate level, and we are following the same strategy here. With nesting there are 3 criteria for evaluating any possible nesting structure:

1. Are Lambda’s much different than 1?
2. 2 * (LL1 - LL2) is Chi-Squared with DF the difference in parameters
3. Does the nesting make sense?

When we apply a nesting structure and get lambdas near 1, then we probably don’t need the nesting. In general, we prefer lambda values less than .9.

The second criterion is the official criterion for testing statistical significance. Adding nests should improve the log likelihood (since lambda = 1 means no nest). Two times the improvement in log likelihood is chi-squared where the degrees of freedom are the number of additional lambda parameters. As is the case with other significance tests, more sample is likely to make things more significant.

The third criterion is just a reminder of using common sense, and expert knowledge to the extent that it is available.

In our case study, with 57 SKUs, we built 22 nests at the bottom level. Note that 3 items did not go into a nest. Or more exactly each of the 3 items went into its own 1 item nest. The aggregate model showed that there is quite a bit of correlation among the alternatives. With one exception, the lambda parameters are all much lower than 1.

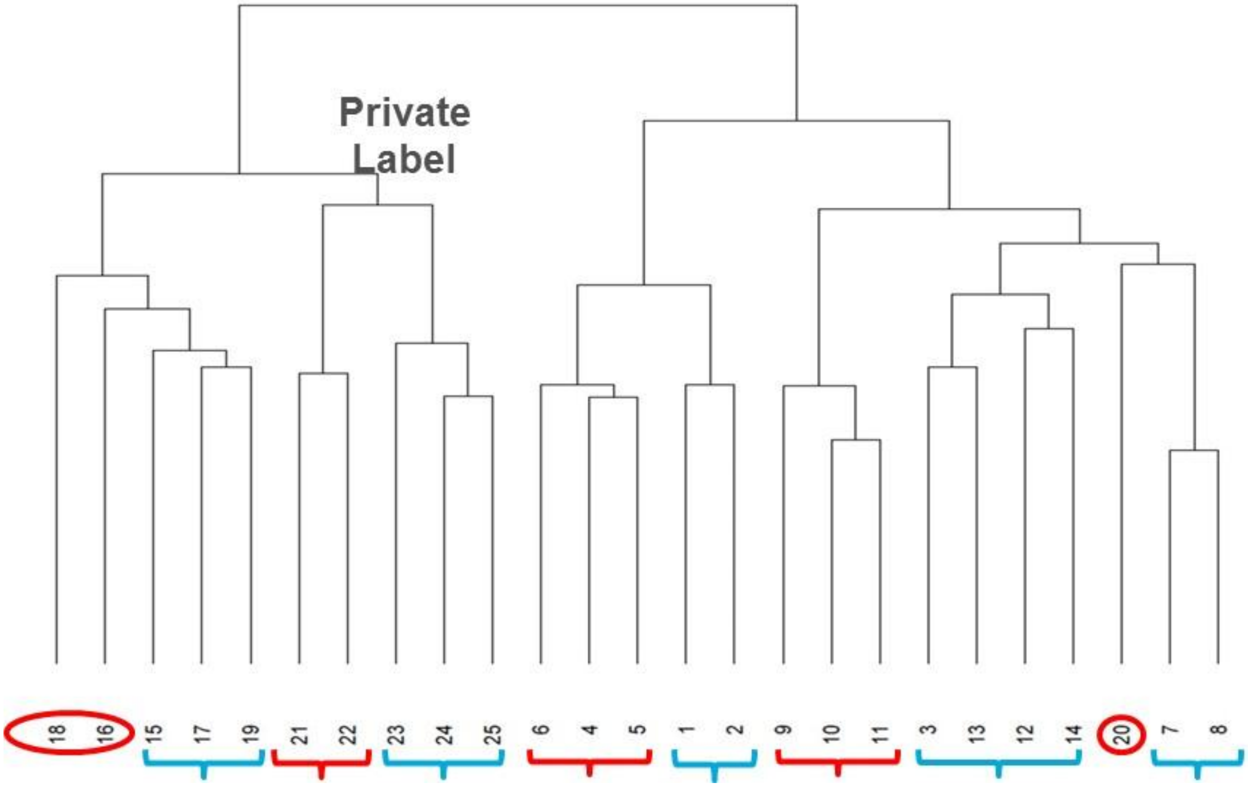
| | | | | | | |
|---|---|---|---|---|---|---|
| **0.516** | **0.632** | **0.363** | **0.380** | **0.259** | **0.753** | **0.760** |
| **0.768** | **0.570** | **0.489** | **0.643** | **0.340** | **0.871** | **0.910** |
| **0.618** | **0.692** | **0.671** | **0.723** | **0.780** | **0.780** | **0.739** |
| | | | | | | **0.876** |

In general the more items that go into a nest, the more likely the corresponding lambda will go toward 1. Part of the reason for the strong nesting here is that we have 2–4 items per nest.

## 3.4 Continue Building the Nest Upward

Having completed the bottom level of nesting, we can move up to the next level. Of course, the next level could simply consist of all 22 nests going into the final selection. But it is better to repeat the same testing at the next level up, combining some of these 22 bottom level nests.

The process is exactly the same. We first compute the overlap matrix. Only this time we use the nests from the lower level. So given two nests A and B, we compute the number of respondents who chose an item in A&B divided by the number of respondents who chose an item in A or B. We use all 22 multiple item nests plus the 3 nests with only 1 item. We again run the dendrogram:



The brackets show the 2nd level nesting that we used. It is worth noting that nests 21–25 consisted of Private Label brands. The SKUs to the left were branded but more economical. While on the right hand side we had higher quality premium brands.

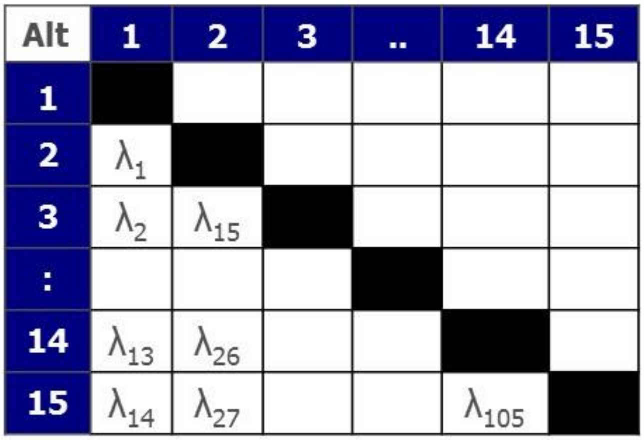
This kind of nesting structure can be shared with the client, and provide additional insight into which items compete most strongly with each other.

### 3.5 Final Comments on Nest Building

I want to conclude this section with a few comments about nest building.

One crucial point is that an item can be in more than one nest. While we have not shown that in our case study, there is no mathematical requirement that an alternative be in just one nest. If an alternative is in more than one nest then we need to sum the probabilities for that alternative. For example, if we have an alternative in 3 different nesting structures, it would appear 3 times at the base level. We would then compute the nests as described and get 3 probabilities for the alternative conditional on each nest.

Paired combinatorial logit is very similar to a covariance matrix, where the elements of the covariance are lambdas for each nest. For example, with 15 items we would have 15(14)/2 = 105 lambdas as follows:



| Alt | 1 | 2 | 3 | .. | 14 | 15 |
|---|---|---|---|---|---|---|
| 1 | ■ | | | | | |
| 2 | $\lambda_1$ | ■ | | | | |
| 3 | $\lambda_2$ | $\lambda_{15}$ | ■ | | | |
| : | | | | ■ | | |
| 14 | $\lambda_{13}$ | $\lambda_{26}$ | | | ■ | |
| 15 | $\lambda_{14}$ | $\lambda_{27}$ | | | $\lambda_{105}$ | ■ |

This is a symmetric matrix with 1's in the diagonal. Computing lots of lambdas will likely result in overfitting. So we suggest something like a structured covariance approach, where each lambda is defined by a general function related to each alternative. See for instance the work on this by Jeffrey Dotson in the context of probit.

## 4.0 Extending Nested Logit Beyond Aggregate Models

So far we have discussed nested logit at the aggregate level. But we know that accounting for respondent heterogeneity significantly improves our conjoint predictions. The two most common ways to do this are with Hierarchical Bayes and Latent Class. In our case study each of the 1,157 respondents evaluated 16 conjoint tasks. For each respondent we removed two tasks from the estimation and used them as holdouts. We ran the following models, and computed the fit for the holdouts:

| | Log Likelihood | % Imp Over Agg |
|---|---|---|
| Aggregate | -7,614.6 | |
| Latent Class (30 Segment) | -5,248.5 | 31.1% |
| Hierarchical Bayes | -5,164.2 | 32.2% |
| Latent Class Ensemble (20)* | -4,708.4 | 38.2% |

The last row shows results for Latent class ensembles, where we ran 20 Latent class solutions, each with 30 segments. For more details on this see Kevin Lattery's paper in the 2015 Sawtooth Proceedings.

Estimating a nested logit at the aggregate level resulted in a log likelihood of -7,524.6. While this was statistically significant, it is only a 1.2% improvement over the simple aggregate model. One of the important points with nested logit models is that we should not expect big improvements in log likelihood. The improvement is in how the shares shift. We apply nested logit to improve sourcing computations, not because they dramatically improve respondent-level fit.

The table above also shows that aggregate models, with or without nesting, are far less accurate than models that account for respondent heterogeneity. In the next sections, we will discuss how to extend nested logit to Latent Class and Hierarchical Bayes.

## 4.1 Extending Nested Logit to Latent Class

The easiest and most direct way to extend nested logit models is using Latent Class. Since Latent Class estimates aggregate-level models using respondent weights, everything from earlier sections applies.

The only real issue is time needed for estimation. Latent Class models involve iterative estimation, updating weights and estimating models again. For a standard logit this is not so bad. But estimating a nested logit takes much longer, especially when we have lots of alternatives and nests. With our case study, the aggregate level model took just over one minute on my personal PC, but the nested logit took 23 minutes. For 30 segments (and using parallel processing on my 4 core machine), it took 7 hours and 30 minutes to compute one Latent class solution.

Because ensembles of Latent Class models perform better, we also ran 20 LC models over the course of one week using two computers. While this produced the best results, clearly one must budget for time. Of course, since each ensemble member is estimated independently we could have rented 20 computers in the cloud to run 20 Latent class solutions in less than 8 hours.

One alternative is to use the lambda parameters from the aggregate model. Since we developed and tested the structure in aggregate, we already have these aggregate level values. Obviously this is a compromise based on time. But it is theoretically justified. In a standard Latent class model we assume all lambda parameters of 1. Here we have estimated better lambda parameters and are using those in our Latent class solution.

## 4.2 Extending Nested Logit to Hierarchical Bayes

At first glance it might seem tempting to estimate nested logit with HB by simply changing the utility function and adding the lambda parameters as additional parameters. One could do this for instance using the R package RSGHB. This however proves to be disastrous. We strongly recommend against doing this. In our case study the fit gets much worse, lowering hold out log likelihood from -5,164.2 to -5,828.2. This in fact makes HB significantly worse than a single Latent Class solution without nesting.

Part of the problem is that the lambda parameters don’t belong in the same covariance matrix as the other betas. While doing separate draws from 2 covariance matrixes would likely help, we still conjecture that we would likely overfit the data. Estimating respondent-level betas in the presence of many alternatives already has significant error, and in fact many are now turning away from respondent-level betas. Adding lots of respondent-level lambdas will significantly exaggerate this problem. We simply believe that respondent-level lambdas are not feasible unless potentially there are only a few lambdas to estimate via a separate covariance matrix. In cases like ours where we are modeling many SKUs with many nests and levels, respondent-level lambdas will not work.

However, we can borrow our previous idea and use global lambdas. This can be done in two different ways. First, we can fix the lambdas to be the values discovered during the aggregate level modeling and testing. We then change the utility function and estimate betas only, assuming the fixed lambdas.

A second way is to estimate the lambdas as fixed effects during the HB estimation. Details of the process can be found in section 12.7.3 of Kenneth Train’s book *Discrete Choice Methods* *with Simulation*. The R package RSGHB allows one to specify both fixed (global) and random (respondent-level) parameters.

We show the results of both methods in our table below, along with the Latent Class versions.

| | No Nest | Aggregate Lambdas | Lambda as Fix Effect | Latent Class Lambda |
|---|---|---|---|---|
| Aggregate | -7,614.6 | -7,524.7 | | |
| Hierarchical Bayes | -5,164.2 | -5,160.3 | -5,091.4 | |
| LC (30 Segment) | -5,248.5 | -5,175.4 | | -5,142.9 |
| LC Ensemble (20) | -4,708.4 | -4,667.7 | | -4,656.3 |

As noted before, the Latent Class ensemble with 20 LC models of 30 segments performed best, but took a week to estimate. In all cases, using the aggregate-level lambdas performed almost as well. It also gave us more predictable changes in sourcing that better aligned with our expectations. Unfortunately real market shares or even aggregate shares in the conjoint were not available.

When we estimated lambdas (rather than using the aggregate level) we found significantly flatter lambdas. For the HB model with a fixed effect, we only had 11 lambdas (out of 37 total) that were less than .9, and only 3 less than .8. We also found that it gave very similar predictions to the standard HB model without nesting. The HB model where we estimate only betas assuming the aggregate-level lambdas gave us better sourcing effects.

At this point we are not convinced that estimating lambdas beyond the aggregate level is worth the extra effort. Fixing the lambdas at their aggregate level is better than simply assuming all of them are 1 which we usually do. In addition, using aggregate-level lambdas also makes it an easier story to communicate.

## 5.0 Conclusion

For most models the assumption of IIA works fine, especially when we incorporate respondent heterogeneity with methods such as Latent Class or Hierarchical Bayes. But in some cases, violations of IIA at the sub-aggregate level trickle up and create noticeable problems with sourcing. In those cases, a simple solution is to do post-hoc fixes in the simulation. We can add correlation on the back-end. But that means we are modeling the data one way, and simulating a different way. Here we have shown how to bring the correlation of alternatives into our model.

Nested logit is only one such approach, and we are open to other ideas. It is a classical workhorse that we have shown how to extend to Latent Class and HB models. There may be other models that capture sourcing even better. Nested logit is fundamentally an aggregate approach that extends rather easily to Latent Class. But it is not a natural fit with HB. Other approaches developed with HB in mind may work much better.

In the meantime, our paper shows how to solve sourcing problems when they arise. We have seen this problem especially in cases where there are many SKUs on a shelf set. We showed how to develop and test the nesting structure. Again, we have revealed our approach using pairwise overlap matrices. But this is not the only method. It is a diagnostic tool that we have found works well in many cases. Nested logit is both a science of testing and an art of knowing what structures to test. That art can be guided but it is ultimately about exploring a subset of options based on data and human skill. We may not find the absolute best nesting structure, but that's okay. We can develop a nesting structure that improves our model, makes sense, and is validated with statistical testing.

Kevin Lattery

# Mining and Organizing User-Generated Content to Identify Attributes and Attribute Levels

**Artem Timoshenko**
**John R. Hauser**
*MIT Sloan School of Management*[38]

## Abstract

We investigate User-Generated Content (UGC) as a source of customer needs from which to identify attributes and attribute levels for a high-craft conjoint analysis study. Non-informative and repetitive content crowd out information about customer needs in a large corpus of UGC. We design a machine-learning hybrid approach to enhance customer-need extraction making it more effective and efficient. We use a convolutional neural network (CNN) to identify informative content. Using pre-trained word embeddings, we create numerical sentence representations to capture the semantic meaning of UGC sentences. We cluster sentence representations and sample sentences from different clusters to enhance the diversity of the content selected for manual review. The final extraction of customer needs from informative diverse sentences relies on human effort. In a proof-of-concept application to oral care, we compare customer needs identified from UGC to customer needs identified from experiential interviews. First, our analyses suggest that, for comparable human effort, UGC allows identifying a comparable set of customer needs. Second, machine learning enables analysts to identify the same number of customer needs with less effort.

This paper summarizes results from Timoshenko and Hauser (2017). All copyrights remain with the original paper, which provides much greater detail. Non-exclusive permission is given to Sawtooth Software to publish this paper.

## Motivation

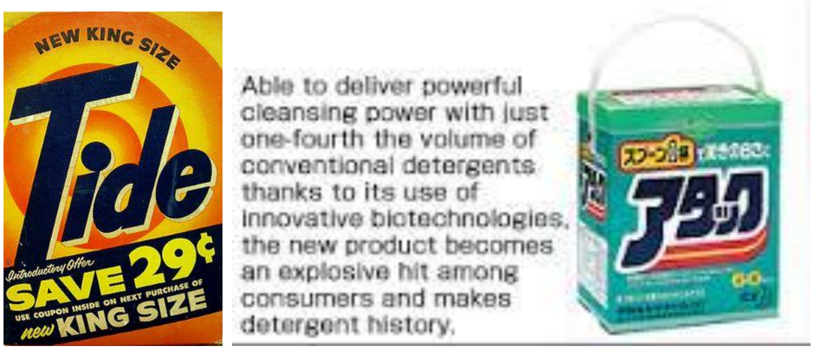
A conjoint analysis study is only as good as the attributes upon which the study is based. Missing important attributes lowers the quality of the study and leads to inefficient product development. Identifying new highly-valued attributes and attribute levels leads to major breakthroughs in product strategy. Consider "Attack," a laundry detergent introduced by the Kao Group in the 1980s. At the time, the customer needs for laundry detergents were well established: cleaning, safe and gentle, good for the environment, ready to wear after drying, easy to use, smell fresh and clean, and value. To design new detergents, most manufacturers focused on combining attributes to address these customer needs. Perceived "value" played a major role in the market for detergents. For example, detergents were sold in large "high-value" boxes to enhance perceived value. Figure 1 compares a vintage Tide box with Attack's packaging at its launch.

Kao did not limit itself to established attributes and attribute levels. Japanese consumers did not have the space to store laundry detergent in their apartments and, as a result, they went to the

[38] MIT Sloan School of Management, Massachusetts Institute of Technology

store often. Consumers commonly went by bicycle or by foot. Kao recognized an unmet customer need and the corresponding attribute level (the need for small package for the same cleaning power). Kao launched Attack, a highly-concentrated detergent in an easy-to-store and easy-to-carry package. Laundry customers were willing to pay a substantial price premium for this product and, within a year, despite the higher price, Attack commanded almost 50% of the Japanese laundry market (Kao Group 2016). Other firms, including US-based firms, were slow to identify this customer need and did not immediately include the "low-package-size" attribute level in their marketing studies, which gave Kao a competitive advantage.

**Figure 1. Vintage Tide Detergent Box and Attack's Package at Launch**[39]



Able to deliver powerful cleansing power with just one-fourth the volume of conventional detergents thanks to its use of innovative biotechnologies, the new product becomes an explosive hit among consumers and makes detergent history.

Examples of successful major innovations based on newly identified attributes and underlying customer needs include the touchscreen features in the smartphone category and Procter & Gamble's Swiffer mop (Continuum 2016). Even point-of-care blood-gas testing in intensive care units of hospitals was revolutionized when the need for new attributes for these important medical instruments was recognized, analyzed, and satisfied.

These examples come from product development, but conjoint analysis is also used widely to value patents and copyrights (Cameron, Cragg, and McFadden 2013). Accepted litigation practice pairs a marketing expert, who provides estimates of willingness to pay, with a "damages" expert, who handles the implications of WTP. The damages expert testifies about the value of the patent or copyright. Recently, Allenby *et al.* (2014) proposed that the marketing expert play both roles. Instead of computing WTP, the authors propose that conjoint analysis be used directly to estimate the change in market price that is due to the patent. They propose that a conjoint analysis simulator be used to determine the (Nash) market equilibrium prices at which all firms in the market simultaneously select maximum-profit prices, each assuming the other firms do not change their prices. Their proposed method requires a reasonably complete set of attributes, because equilibrium prices depend upon the error term in conjoint analysis which, in turn, depends on unmodeled attributes. See Eggers, Hauser, and Selove (2016) in this volume. The courts have intuited this dependence. When conjoint analysis is used for more than WTP (or willingness to buy, WTB), some courts have disallowed testimony from conjoint analysis experts because the courts perceive that the attribute description is inadequate (e.g., Alsup 2012).

Whether a conjoint analysis is used to price a product, identify new product opportunities, estimate the impact of a change in attributes, or value copyrights and patents, it is important that the conjoint analysis study is based on a rich set of attributes for the product category. The

[39] Tide image from https://www.pinterest.com/blacklab3/vintage-soap/. Attack image from http://www.kao.com/group/en/group/history_01.html.

accuracy and relevance of the conjoint analysis study depends on the quality and completeness of the attribute-based description.

## Typical Approaches to Identify Attributes

### Direct Approaches

Often, the client provides a list of attributes and attribute levels and asks the analyst to design and execute the conjoint analysis experiment. This is a perfectly fine approach, but pushes the responsibility back to the client to specify an appropriate list of attributes. Alternatively, an analyst might search competitive websites, search websites that compare and contrast products, and search websites that make recommendations. Advertising claims complement these Internet searches. If the market is relatively stable, or if the conjoint analysis is used for WTP or WTB, then well-conducted Internet searches are an efficient means to identify the attributes for the conjoint analysis study. Internet searches are less useful if the market is in flux, or if the goal is to identify new innovations. “Unarticulated” needs and attributes might not be found in these Internet searches because no existing product has the attributes. New opportunities could be missed. Analysts must also be careful because comparison websites focus on points of difference among products. They might miss basic “must have” attributes.

### Indirect Customer-Based Approaches

Indirect customer-based approaches begin directly with the customer. Focus groups and experiential interviews enable customers to articulate their needs and desires for the product category. The analyst experiences the experiences of the customers. Rather than asking directly about attributes, the analyst seeks first to understand the customers’ needs and then translates those customer needs into attributes (solutions) that address the customer’s expressed needs. Fortunately, there are a variety of proven methods to translate customer needs into attributes, including hedonic regression, Quality Function Deployment, and the Brunswik “lens” model (Brunswik 1952, Hauser and Clausing 1988, Sullivan 1986).

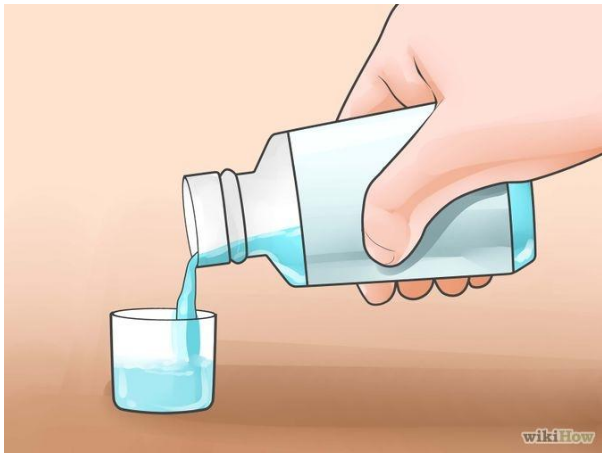
The direct approaches are easier to implement and less expensive, but the indirect customer-need-based approaches provide certain advantages. Indirect approaches identify a broad set of attributes with less functional overlap. This is particularly valuable because survey formats and respondent attention often limit the number of attributes and attribute levels. Furthermore, indirect approaches often identify unmet customer needs that lead to successful innovations.

Our study focuses on identifying customer needs for an indirect approach. We rely on established methods to translate customer needs into attributes and attribute levels.

### Customer Needs versus Customer Solutions

Customer needs, as used in this paper, are abstract statements that describe what a customer seeks to obtain from a product in the category. For example, in oral care, a customer need might be: “Easy to know the correct amount of mouthwash to use.” Customer needs are purposefully abstract so that they provide sufficient flexibility for the firm to design attributes that fulfill customer needs. With these definitions, attributes in conjoint analysis are solutions to customer needs. For example, a solution to the customer need might be to put “ticks on a cap that is used for dosage” or “pictures and numbers on the bottle to indicate dosage.” See Figure 2.

**Figure 2. Attribute-based Solution to Customer Need to Know Easily the Correct Amount of Mouthwash to Use**



## Voice of the Customer

A structured set of customer needs is often called the “voice of the customer (VOC).” The most common VOC method consists of four steps: (1) experiential interviews with customers, (2) sentences highlighted by multiple human judges, (3) “winnowing” to obtain a reduced, non-redundant set of customer needs, and (4) methods to organize the customer needs into an hierarchy of “primary,” “secondary,” and “tertiary” customer needs (Ulrich and Eppinger 2016; Griffin and Hauser 1993; Herrmann, Huber, and Braunstein 2000). There are at least two common procedures to organize customer needs into an hierarchy: (1) affinity groups where customers, themselves, sort the needs, and (2) card-sort methods where customers sort together customer needs that are similar and analysts cluster customer-need co-occurrence matrices. Figure 3 provides an example of the first two levels of a customer-need hierarchy that was delivered to an oral-care client. This VOC was produced by a marketing consulting firm with almost thirty years of experience in the voice of the customer.

**Figure 3. Voice of the Customer for Oral Care Products**

| | |
|---|---|
| FEEL CLEAN AND FRESH (SENSORY) | Clean Feeling in My Mouth<br>Fresh Breath All Day Long<br>Pleasant Taste and Texture |
| STRONG TEETH AND GUMS | Prevent Gingivitis<br>Able to Protect My Teeth<br>Whiter Teeth |
| PRODUCT EFFICACY | Effectively Clean Hard to Reach Areas<br>Gentle Oral Care Products<br>Oral Care Products that Last<br>Tools are Easy to Maneuver and Manipulate |
| KNOWLEDGE AND CONFIDENCE | Knowledge of Proper Techniques<br>Long Term Oral Care Health<br>Motivation for Good Check-Ups<br>Able to Differentiate Products |
| CONVENIENCE | Efficient Oral Care Routine (Effective, Hassle-Free and Quick)<br>Oral Care “Away From the Bathroom” |
| SHOPPING / PRODUCT CHOICE | Faith in the Products<br>Provides a Good Deal<br>Effective Storage<br>Environmentally Friendly Products<br>Easy to Shop for Oral Care Items<br>Product Aesthetics |

## User Generated Content (UGC)

User-generated content (UGC) is text (and pictorial) content about products that customers themselves generate. For example, Twitter posts, customer blogs, and customer reviews are all UGC. UGC might also include customer complaint data or data collected from customer-help records. UGC is an exciting new source of information from which customer needs (and conjoint analysis attributes) can be extracted. UGC is often available quickly and at low incremental cost to the firm. UGC is updated automatically and never gets stale.

However, UGC presents its own challenges. First, there are often too much data for human readers to process. For example, there are over 115,000 oral-care reviews on Amazon consisting of over 400,000 sentences. Human readers just cannot process that entire corpus. Second, much of the data in UGC are repetitive and not relevant. Sentences such as “I recommend Crest for oral care” does not express any customer need. We expect, and our analysis confirms, that most of the UGC on oral care concentrates on a relatively few needs. Third, UGC data are unstructured and mostly-text based. Identifying customer needs requires a thorough understanding of the content, and the unstructured nature of UGC complicates automatic analysis.

## Our Goals

### UGC versus Experiential Interviews

Our first goal is to compare in completeness and quality a set of customer needs, identified from UGC, to customer needs identified by standard methods as practiced by experienced analysts working from high-quality experiential interviews. Ideally, UGC-based customer needs should (1) have a substantial overlap with interview-based customer needs, (2) miss relatively few interview-based customer needs when limited to comparable analyst effort, and (3) include customer needs that were not identified from an exhaustive search of experiential-interview transcripts. We feel that if we confirm that customer needs from UGC satisfy these criteria then we validate UGC as a viable replacement for costly experiential interviews.

### Machine-Human Hybrid versus Human-Only Processing

Our second goal is to use machine learning (deep learning) to streamline the identification of customer needs from UGC. In particular, we seek to use machine learning to eliminate non-relevant content and organize the remaining content to minimize redundancy.

For example, suppose that analysts, who are experienced in the use of VOC methods, have the capability of reading $N$ sentences from UGC to identify customer needs. (Their capability might be limited by time, monetary budgets, or simply attention.) Not all $N$ sentences will be relevant and many sentences will describe redundant customer needs. Let's suppose that the analysts can identify $K_o$ unique customer needs. A machine-human approach is more efficient if it can identify at least $K_o$ customer needs with human effort that is less than or equal to that which would have been required for VOC experts to review $N$ random sentences and identify $K_o$ customer needs. (Computational costs are trivial compared to human effort.)

If we demonstrate that the machine-human hybrid is more efficient, then with continuous improvement through application the evolved method might be able to optimize the machine-human hybrid and achieve the best human-effort cost per identified customer need. (We assume that the machine-learning method is fully programmed. The computation cost is a very small fraction of human effort.)

### Optimization of Human Effort

In our scheme, there are multiple types of human effort that enter any analysis. In standard VOC methods, experiential interviews are extremely costly. UGC eliminates recruiting, interviewing, and transcription costs. In the machine-human hybrid method, there are two types of human effort required. Human analysts review sentences to determine whether the sentences are "informative" or "non-informative." Then, human analysts review informative sentences to extract customer needs. The former is less onerous and time-consuming than the latter. For the purposes of this paper, we leave the optimization of human effort to future research. Such optimization requires that we quantify the value of additional customer needs and quantify the effort costs of interviewing, customer-need extraction, and informative vs. non-informative classification.

## A Proposed Machine-Human Hybrid for Attribute Identification

### Why a Machine-Human Hybrid

When machine-learning methods improve, we might be able to automate all stages in the identification of customer needs from UGC. To date, the final stage has defied automation. Formulations of customer needs must be precise for subsequent analyses. Moreover, the machine-learning methods are not sufficiently sensitive to semantic context to extract abstract customer needs from informative content. UGC is unstructured and not necessarily generated to articulate customer needs. Context matters and customer needs appear to be more than "buckets of words." For example, bucket-of-word methods, such as Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan 2003) and LDA with hidden Markov models (LDA-HMM) (Griffiths *et al.* 2004) do not seem to capture the semantic context necessary for identifying customer needs. But stay tuned.

We have successfully automated two critical tasks in the analysis of UGC: identifying informative content and sampling a representative and diverse set of content for review. The resulting machine-human hybrid is more efficient, and equally as effective, as a pure human-effort-based method. We feel this is substantial progress in a relatively short time. Analysts have had almost thirty years of continuous improvement to optimize human-effort-based VOC methods.[40] VOC identification by experienced analysts is a challenging benchmark.

### Overview of the Machine-Human Hybrid

Table 1 provides an overview of the four stages in our proposed method. The stages are:

1. ***UGC***. Rather than relying on expensive experiential interviews, we harvest readily available UGC from either public sources or propriety company databases.
2. ***Identify Informative Content***. We use a machine-learning classifier called a convolutional neural network (CNN) to filter out non-informative sentences so that the remaining corpus is rich in informative content. Because a CNN is a supervised learning method, it must be "trained." Training requires human effort to classify a subset of sentences as informative vs. non-informative. In practice, the number of training sentences should be a small fraction of the corpus.
3. ***Sample Diverse Content***. We cluster "sentence representations" to select a set of sentences likely to represent diverse customer needs. Sentence representations are, in turn, based on dense numerical representations of words that capture semantic meanings.
4. ***Final Extraction of Representative Customer Needs***. Analysts review the winnowed, informative sentences to identify customer needs. In the machine-human hybrid approach, this final stage is based on human effort and is the same task as that used in existing human-effort-based methods.

We now describe the two machine-learning methods that we customized to the identification of customer needs. We then describe a proof-of-concept application to oral care.

[40] Consulting firms, with experience in VOC methods, make human effort more efficient with software that makes it easy to highlight phrases in interview transcripts. Additional "bookkeeping-like" software makes it easy to keep track of redundant phrases during the winnowing process. Such proprietary software does not have the capabilities to be called machine-learning. These firms have also optimized human effort through training and experience.

Table 1. Automating Attribute Identification—Machine-Human Hybrid

| Traditional | Machine-learning Hybrid |
|---|---|
| Experiential interviews | User Generated Content |
| Highlight informative sentences manually | Machine learning (convolutional neural network, CNN) identifies informative sentences |
| Reduce customer-need redundancy manually (winnowing) | Cluster numerical “sentence representations” to remove sentence redundancy and thus identify diverse customer needs |
| Extract customer needs manually from interview-based sentences | Extract customer needs manually from informative diverse UGC sentences |

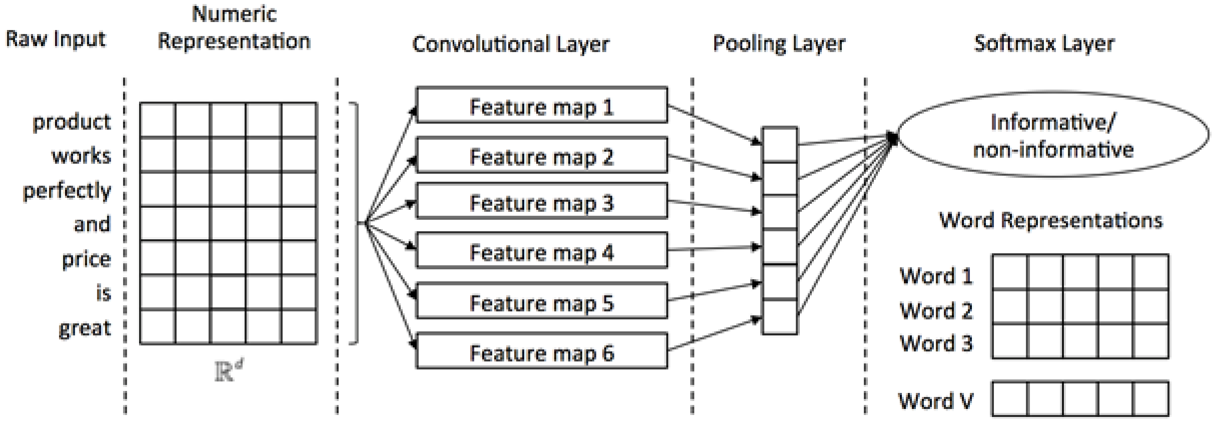
## PREPROCESSING UGC TO IDENTIFY SENTENCES WITHIN THE UGC

Sentences are most likely to contain customer needs and are a natural unit by which human analysts process either experiential interviews or UGC. But in UGC, customers do not always use a sentence structure. We preprocess raw UGC to transform the UGC corpus into a set of sentences. We use an unsupervised sentence tokenizer from the natural language toolkit (Kiss and Strunk 2006). We automatically eliminate stop-words (e.g., “the” and “and”) and non-alphanumeric symbols (e.g., question marks and apostrophes). We transform numbers into number signs and letters to lower case. We further screen sentences to account for the artifacts of grammatical or punctuation errors in UGC. In particular, we drop sentences that are too short (less than three words) or too long (more than ten words). UGC tends to have many fewer compound sentences than experiential-interview transcripts.

## CONVOLUTIONAL NEURAL NETWORK (CNN)

We use a convolutional neural network (CNN) on the corpus of sentences after preprocessing to classify sentences as either informative or non-informative. A CNN is a supervised classification model (e.g., Kim 2014). We use a CNN to transform numerical representations of sentences into a prediction of whether or not the sentence is informative. A CNN has multiple types of layers and can have multiple layers of each type. Figure 4 illustrates the types of layers that are contained in our CNN. (Our CNN is not an off-the-shelf CNN, but rather customized for our application.) The two key properties of the CNN are that (1) the CNN learns how to quantify and classify sentences simultaneously in the model, and (2) the model is able to process input (sentences) of different length.

**Figure 4. Examples of the Types of Layers in our Convolutional Neural Network**



### Numeric Representations of Words

For every word in the English-language dictionary, the CNN represents the word by a numerical vector. We use pre-trained 300-dimensional word embeddings as described in the next section. If the word embeddings were unavailable, and with sufficient training data, a CNN could be used to learn word representations simultaneously with other parameters. The CNN quantifies the sentence by concatenating the representations of the words.

### Convolutional Layers

A convolutional layer begins by applying filters to the sentence representation. A filter selects varying contiguous subsets of the sentence representations and weights the elements of the subset. The CNN then applies non-linear transformations, such as a logistic function, to the weighted subsets. The result of the application of this transformation to various parts of the sentence representation is called a “feature map.”

We calibrate the weights used in the filters by training the CNN on the sentences that have been coded by human effort. The number of filters and their sizes are hyperparameters of the CNN. We select these hyperparameters before the CNN is trained. We tune the hyperparameters with cross-validation.

### Pooling Layers

Convolutional layers often require many parameters and can become too complex to calibrate. If multiple convolutional layers are stacked without any dimensional reduction, then the number of parameters explodes. (The number of parameters also explodes if there are too many feature maps.) To maintain a feasible number of parameters, CNNs use pooling layers, which transform feature maps into shorter vectors. We use a $max$-pooling-over-time layer in which we retain the largest feature from each feature map produced by a convolutional layer (Collobert *et al.* 2011).

### Softmax Layer

The final layer, called a softmax layer, in the CNN transforms the output of the final pooling layer into a prediction of whether the sentence is informative ($y = 1$) or not informative ($y = 0$). The softmax layer is a binary logit model applied to the output of the last pooling layer. The parameters of the logit model are calibrated with the training data. In our application, we assign a sentence as informative if the estimated probability is greater than 50%. Future applications might assign sentences to categories for further review based on other criteria.

### Number of Each Type of Layer

In our study, we stacked 3 convolutional layers and 1 pooling layer to generate input for the softmax layer. Each convolutional layer generates 40 feature maps. Performance of the trained CNN depends on a particular combination of layers and on the number of feature maps in convolutional layers. We used cross-validation to select these characteristics of the model.

### CNNs vs. SVMs

Readers may be familiar with the use of support-vector machines (SVMs) for classification. CNNs have an advantage relative to SVMs because CNNs automatically and endogenously identify feature maps. In contrast, an SVM depends critically of the quality of the features used in the SVM. SVM features are often handcrafted, specific to application, dependent on context, and require substantial human effort. CNNs provide comparable performance to handcrafted SVMs without this substantial application-specific human effort (Kim 2014).

## Clustering Sentence Representations

Armed with a corpus of informative sentences, we use machine learning to reduce redundancy. We cluster sentences that have similar semantic meaning and then sample from each cluster in proportion to the size of the cluster. For a given number of sentences, redundancy-reduced sentences are more likely to contain diverse needs than a random sample of informative sentences. Because the clustered corpus is designed for maximum diversity, it is more likely (for a given $N$) to yield a complete set of customer needs.

In order to cluster sentences, we create numerical representations of the sentences that capture semantic meaning. The transformation for clustering is different than the concatenation for CNN classification, but both transformations are based on machine-language constructs known as “word embeddings.” We first describe word embeddings and then describe how we aggregate word embeddings to sentence representations.

### Word Embeddings

Word embeddings are the numeric vectors that capture the semantic meaning of words. The basic concept is that semantically similar words appear in similar contexts. Information about the contexts is then used to represent words in the numerical space. We rely on a high-quality pre-trained set of word embeddings that have remarkable properties. For example, if a word embedding, $v(w_i)$, is a vector representation of word $w_i$, then the $v(w_i)$ have the following properties (Mikolov *et al.* 2013a):

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen})$$

$$v(\text{walking}) - v(\text{swimming}) + v(\text{swam}) \approx v(\text{walked})$$

$$v(\text{Paris}) - v(\text{France}) + v(\text{Italy}) \approx v(\text{Rome})$$

We use 300-dimensional word embeddings that were pre-trained on the Google News Corpus using the "Skip-gram" model (Mikolov *et al.* 2013b). The Skip-gram model trains word embeddings by maximizing the average log-likelihood of words appearing within $c$ words of one another in a sequence. For our purposes we simply adopt the word embeddings without further transformation.

### Sentence Representations

In the CNN we concatenated word embeddings. This operation matches the use of filters in the feature maps. To create sentence representations for clustering we use an operation that retains the centrality of the semantic meaning. For our proof-of-concept application in oral care, we adopt the averaging method advocated by Iyyer *et al.* (2015). This operation is based on machine-learning experience. For example, Iyyer *et al.* demonstrate that the average of word embeddings is as effective as explicitly modeling semantic and syntactic structure with neural networks or training sentence representations simultaneously with word embeddings (Le and Mikolov 2014; Tai, Socher, and Manning 2015).

### Clustering Sentence Representations

Because sentence representations have the property that similar vectors represent sentences with similar semantic meanings, we cluster the sentence representations based on the Euclidean-distance norm. To be consistent with the hierarchical structures used in established VOC methods, we use an hierarchical clustering algorithm. Griffin and Hauser (1993) suggest Ward's method, which we adopt. Not only has Ward's method become standard practice for analyzing co-occurrence data but, by using Ward's method, we maintain comparability with the human-effort-based benchmarks that we compare to the machine-human hybrid approach.

### Final Extraction of Customer Needs

The clustered sentence representations, sampled proportional to size, provide a set of informative sentences that are designed to be rich in diverse customer needs. The final stage relies on trained analysts to read each sentence and extract the customer needs. We expect that human-effort extraction is more efficient with informative, diverse sentences than with sentences sampled randomly from the UGC corpus.

## Oral Care Proof-of-Concept, Evaluation, and Comparison to Established Methods

We have three goals.

- Demonstrate that the machine-learning hybrid is feasible and that it can generate a set of customer needs from which attributes can be identified.
- Compare the relative customer-need content of UGC and experiential interviews.
- Evaluate the efficiency of the machine-human hybrid vs. a human-effort-based approach.

We select the oral care category because oral care is best described by a relatively broad and challenging set of customer needs, but the set of tertiary customer needs in oral care is not too large to make the analysis unwieldy.

### "Gold Standard" Human-Based Approach

A professional marketing consulting firm shared with us a VOC that they had delivered successfully to a client. Review Figure 3. The VOC was based on experiential interviews, with sentences highlighted by human analysts aided by the firm's proprietary software. After winnowing, customer needs were clustered by an affinity group. The output was six primary customer needs and 22 secondary customer needs (Figure 3), as well as further elaboration into 86 tertiary customer needs.

### UGC Data

We consider 115,099 oral-care reviews from Amazon.com spanning the period from 1996 to 2014. Preprocessing with the sentence tokenizer produced 408,375 sentences.

### Unique Dataset

To compare the customer-need information in UGC to the customer-need information in experiential interviews, we randomly selected 8,000 sentences from the UGC corpus. The sentence structure of UGC differs from that in experiential interviews. UGC sentences tend to be shorter and less compound. In experiential interviews, sentences tend to ramble as they do in normal conversation. They are not always complete, but make sense in context. Also, the questions asked by interviewers are part of the give-and-take and cannot be ignored. To affect a valid comparison, we asked analysts, with experience extracting needs from interview transcripts, to estimate the number of UGC sentences that would be comparable to those contained in a typical VOC study. They judged the human effort involved in extracting customer needs from 8,000 UGC sentences would be comparable, but slightly less than, the effort involved in extracting customer needs from interview transcripts.

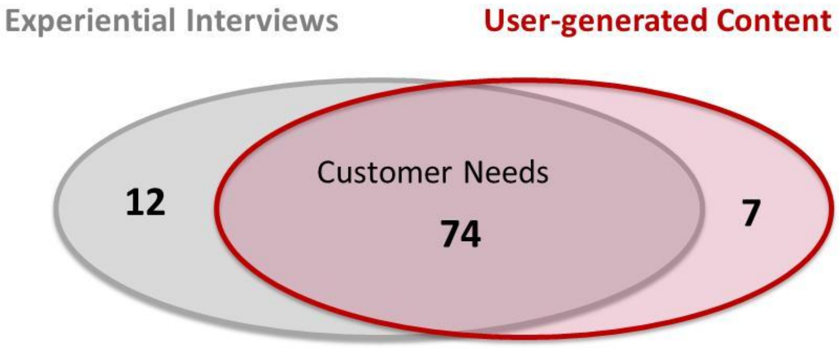
The analysts, who extracted needs from the UGC, were drawn from the same marketing consulting firm that produced Figure 3. This enabled us to maintain a common level of training and experience. For each sentence, the analysts identified all customer needs in the sentence and coded those customer needs against the primary, secondary, and tertiary customer needs in the gold standard. If a tertiary customer need was not in the gold standard, the analysts attempted to assign the customer need to an existing secondary-customer-need group. If the tertiary customer need could not be assigned to a pre-existing customer-need group, the tertiary customer need was given a new number. This data set is unique because the analysts coded *all* customer needs in every sentence of the UGC. Typical practice does not maintain such a map between the source of each customer need and the customer need.

### Information Contained in UGC versus Experiential Interviews

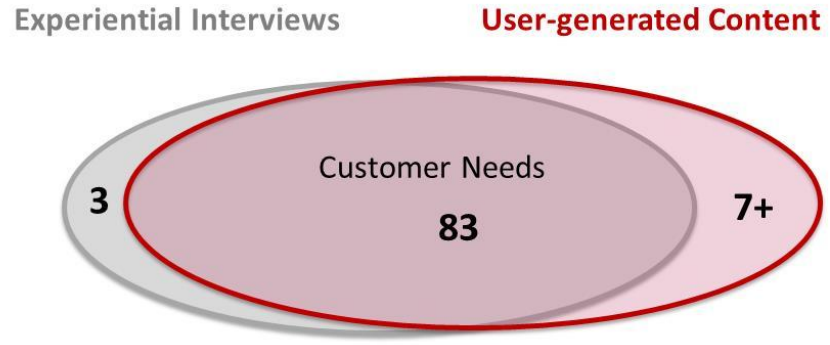
We compared the information contained in the two sources of customer needs. This comparison is summarized in Figure 5a. Of the 86 tertiary customer needs extracted by human effort applied to the transcripts, 74 customer needs (86%) were extracted by human effort from the UGC. Importantly, analysts extracted seven new customer needs from the UGC, customer needs that were not extracted from the experiential interviews. This is impressive. We then asked analysts to examine an additional 4,000 randomly-selected UGC sentences to see if the customer needs that were identified from experiential interviews could be identified from additional UGC. Nine of the remaining twelve needs were identified. See Figure 5b. The analysts' supplementary task was limited; we do not know if the additional 4,000 sentences contained any additional customer needs. (We plan future research to identify the relative importances of the various customer needs.)

**Figure 5. Comparison of Customer-Need Extraction from a Sample of UGC versus Experiential-Interview Transcripts**

(a) Holding Extraction Costs for UGC to be Less than those for Experiential Interviews.



(b) Allowing Higher Extraction Costs for UGC, but Still Saving Interviewing Costs



We conclude that UGC is at least a comparable source of customer needs as experiential interviews. Because UGC eliminates the substantial effort cost involved in scheduling and implementing qualitative interviews, even with the additional 4,000 sentences, the total human-effort cost is less with the machine-human hybrid approach than with the human-only approach. We'll see later in this paper that machine-learning methods make extracting customer needs more efficient, thus enabling analysts to process a UGC corpus larger than 8,000 sentences for the same effort as was used to process transcripts. Further improvement should increase efficiency even more.

Human-effort coding of the 8,000-sentence UGC corpus suggests that 52% of the UGC sentences are informative about customer needs (contain an identified customer need). There was also high redundancy. Ten percent (10%) of the most-frequently mentioned customer needs were articulated in 54% of the informative sentences. These percentages suggest potential efficiency gains due to the CNN and clustering sentence representations.

## CNN

When the training sample, $X$, is larger, the CNN can classify sentences better. Figure 6 plots the ability of the CNN to classify sentences as a function of $X$. Figure 6 reports results up to 6,000 sentences because preprocessing eliminated 1,394 sentences as too short or too long. This left 6,606 sentences eligible for use in training the CNN.

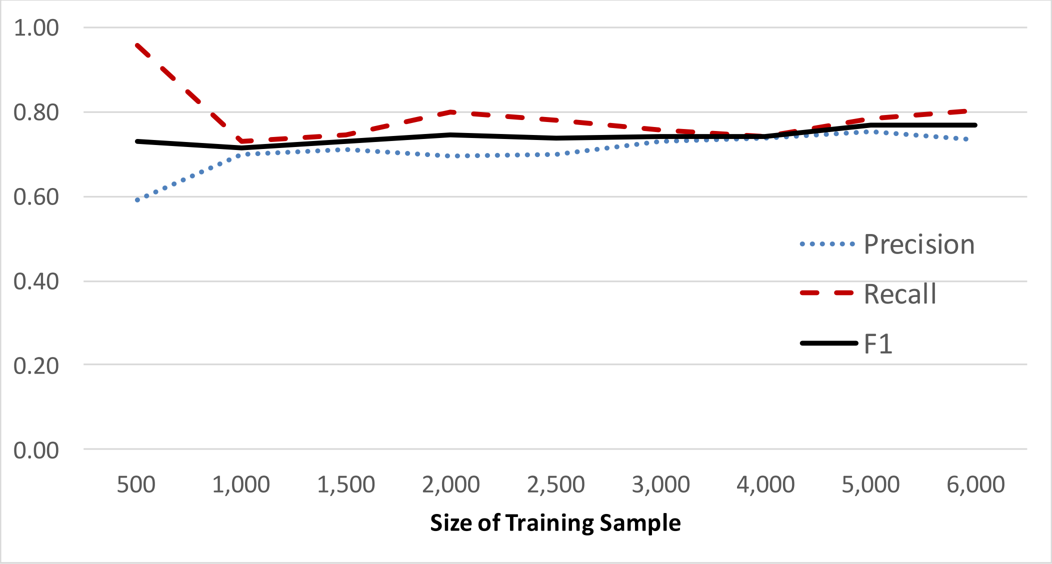
We report three statistics that are common in machine learning. Precision, in machine learning, is comparable to hit rates in conjoint analysis (and not to be confused with the scale factor in conjoint analysis). In sentence classification, precision is the percent of sentences that are informative given that they have been labeled as informative. Recall is the percent of informative sentences that were correctly labeled as informative. $F_1$ is a composite measure equal to:

$$F_1 = \frac{precision \cdot recall}{\frac{1}{2}(precision + recall)}$$

There are tradeoffs in precision and recall as the size of the training sample increases, but their impact on the composite measure, $F_1$, appears to stabilize around $X = 1{,}000$. At $X = 1{,}000$, Figure 6 reports a precision of 70% and a recall of 73%.

The CNN is effective if it identifies customer needs in the UGC corpus that were not in the training data. This was the case. The CNN identified customer needs in the UGC corpus that were not in the training data.

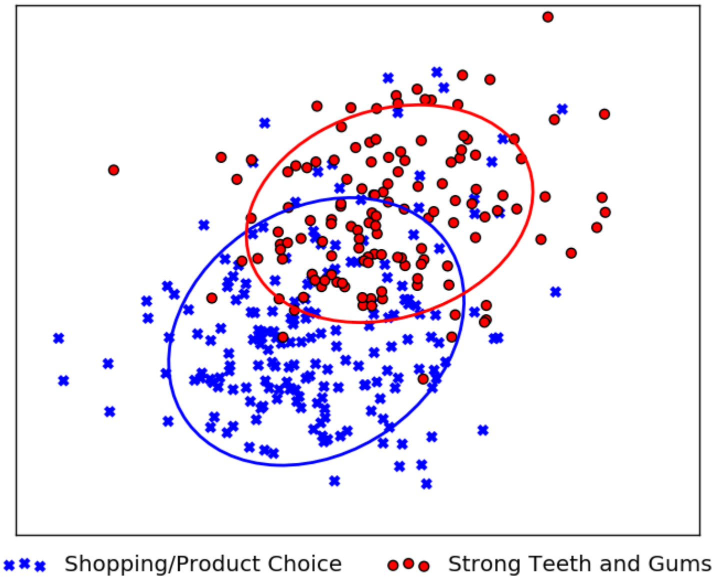
**Figure 6. Precision, Recall, and $F_1$ as a Function of the Size of the Training Sample**



## Clusters of Sentence Representations

To visualize whether or not clustering sentence representations enhanced diversity in customer needs, we use principal components analysis to project the sentence representations onto two dimensions. Information is lost, but we can see visually whether or not customer needs were separated by clustering sentence representations. Figure 7 reports the results.

**Figure 7. Two-Dimensional Projection of 300-Dimensional Sentence Representations**



The red dots are sentence representations that were coded (by human judges) as belonging to the primary customer need of "strong teeth and gums." The blue dots are sentence representations that were coded as "shopping/product choice." The ovals represent the smallest ellipsis inscribing 90% of the corresponding set. Figure 7 suggests that, while not perfect, the clusters of sentence representations did achieve separation among customer needs.

### Gains in Efficiency Due to the Machine-Human Hybrid Method

We use our database to compare counterfactual simulations of the number of customer needs that would have been identified by various methods. We compare the methods for various numbers of sampled UGC sentences. We chose to train the CNN on 5,000 sentences to approximate how we expect the CNN to be used in practice. We believe the larger training sample eliminates randomness in our analysis, but we do not believe that the *relative* comparisons of methods would change.

When we train the CNN on 5,000 sentences, we can hold out 1,606 sentences after preprocessing to eliminate sentences that are too short or too long. At this $X = 5{,}000$, the CNN achieves a precision of 76%, a recall of 78%, and an $F_1$ of 77%. The CNN identifies 1,040 of the 1,606 sentences as informative.

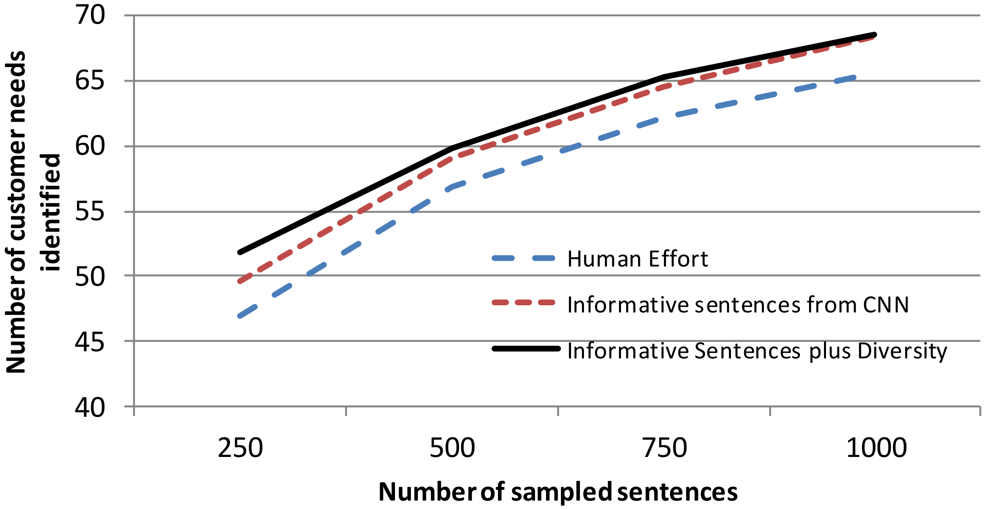
For each of three methods, we compute counterfactuals assuming the analysts have only the resources to review $Y$ sentences for $Y = 250, 500, 750,$ and $1{,}000$. To compare to a human-effort benchmark, we evaluate the customer needs identified from a random selection from the UGC corpus (assuming preprocessing to eliminate sentences that are too short or too long). For example, an analyst would randomly select 250 sentences from the preprocessed corpus and review all 250 sentences. We redraw random samples 1,000 times and average. The results of random selection are shown in Figure 8 by a dashed blue line.

We improve efficiency by using the CNN to identify informative sentences. To test efficiency gains, we randomly select from informative sentences (dotted red line in Figure 8). We increase efficiency further by using the CNN to screen for informative sentences, clustering sentence representations, and selecting from sentence representations proportional to the size of the clusters (solid black line in Figure 8).

Over the range of the counterfactual simulations, Figure 8 suggests that the machine-learning stages enhance efficiency. There are gains due to using the CNN to eliminate non-informative sentences and additional gains due to using sentence representations to seek diversity within the corpus. The gains to diversity decrease with $Y$, but the gains due to the identification of informative sentences continue throughout the range of the counterfactual simulations.

We also interpret Figure 8 horizontally. Human effort requires, on average, 1,000 sentences to identify 65.6 customer needs. If we prescreen with machine learning to select diverse, informative sentences, an analyst can identify, on average, 65.2 customer needs from 750 sentences. These efficiencies represent a human-effort saving of 25%. Given that human-effort-based reviewing of experiential interviews has been optimized over almost thirty years of continuous improvement, these proof-of-concept results are promising. We expect the machine-learning methods, themselves, to be subject to continuous improvement as analysts learn, by trial and error, how best to merge machine learning with human effort.

**Figure 8. Comparison of Efficiencies among Various Means to Select UGC Sentences for Review**



## Discussion and Summary

A high-craft conjoint analysis study requires that attributes and attribute levels be chosen carefully. VOC methods are a proven method by which to identify a complete set of attributes. VOC methods identify customer needs, then established methods, such as QFD, hedonic regression, or the Brunswik lens model, select attributes that are solutions to customer needs.

In this paper we establish that machine-learning methods show promise to extract customer needs more effectively and more efficiently. Machine-learning methods also extract new customer needs that are missed by traditional experiential-interview studies. Once perfected, machine-learning methods applied to UGC will enable conjoint analysis analysts to extract a more complete set of customer needs (attributes) and do so quicker and with less human-effort costs.

### UGC

Our results suggest that UGC can substitute for experiential interviews. In a limited corpus of 8,000 sentences, human analysts were able to extract roughly as many customer needs as would have been extracted from experiential interviews. The overlap was not perfect, but the UGC did identify customer needs not in the interview transcripts. A comparison of Figures 5a and 5b suggests that, with a larger corpus, particularly with efficiencies due to machine learning, UGC should provide sufficient information with which to extract a more-complete set of customer needs than the typical experiential-interview study.

### CNN

The CNN successfully identified non-informative sentences. Future research might optimize the CNN.

## Sentence Representations

Clustering sentence representations increases diversity, especially for small samples. However, as the size of the sample of sentences to review increases, the machine-human hybrid gets close to an exhaustive set of needs and the value of diversity decreases.

## Efficiency gains

Perhaps the largest efficiency gain is the enhanced ability to replace experiential interviews with UGC. Experiential interviews are costly and require calendar time to recruit, schedule, and implement experiential interviews. A typical experiential-interview study requires about 4–5 weeks. UGC can be harvested quickly (less than a day) and at substantially lower cost.

We asked the marketing consulting firm to review 8,000 UGC sentences in depth because they judged that reviewing 8,000 UGC sentences was a conservative estimate of the effort required to review a typical set of experiential interviews. Even with 12,000 UGC sentences, the human effort for extraction is less than the human effort in an experiential-interview study. Both the CNN and clustering sentence representations make the review of the UGC sentences more efficient by as much as 25%. (A percentage we hope to increase with continuous improvement.)

## Machine-Learning Applied to Interview Transcripts

There is nothing to prevent using the CNN and the sentence representation clusters on interview transcripts. We expect to see efficiencies there as well. The machine-human hybrid method applied to the interview transcripts can be useful in product categories where UGC is either not available or not extensive.

## Summary

Understanding customer needs helps define a more complete set of attributes and improves the quality of the conjoint study. Based on our initial proof-of-concept application, we are optimistic about the potential of UGC and machine learning to transform the practice of identifying customer needs. We feel that the CNN and sentence representations are uniquely suited to the analysis of UGC because these methods do more than count words. They look to deep semantic structure as is required in the analysis of UGC.

Artem Timoshenko

John R. Hauser

## References

Allenby, Greg M., Jeff Brazell, John R. Howell, and Peter E. Rossi (2014), "Valuation of Patented Product Features." *Journal of Law and Economics*, 57:3 (August). 629–663.

Alsup, William (2012), "Order Granting in Part and Denying in Part Google's Daubert Motion to Exclude Dr. Cockburn's Third Report," Oracle America, Inc. v. Google, Inc. No. C-10-03561 WHA, United States District Court for the Northern District of California, March 13.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Brunswik, Egon (1952), *The Conceptual Framework of Psychology*, (Chicago, IL: Chicago University Press).

Cameron, Lisa, Michael Cragg, and Daniel McFadden (2013), "The Role of Conjoint Surveys in Reasonable Royalty Cases," *Law360*, October 16.

Collobert, Ronan, Jason Weston, Leon Botto, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011), "Natural Language Processing (Almost) from Scratch," *Journal of Machine Learning Research*, 12, (Aug), 2493–2537.

Continuum (2016). https://www.continuuminnovation.com/en/what-we-do/case-studies/swiffer/.

Eggers, Felix, John R. Hauser, Matthew Selove (2016), "The Effects of Incentive Alignment, Realistic Images, Video Instructions, and Ceteris Paribus Instructions on Willingness to Pay and Price Equilibria," *Proceedings of the 19th Sawtooth Software Conference*, Park City, UT, September 26–20.

Griffin, Abbie (1992), "Evaluating QFD's Use in US Firms as a Process for Developing Products," *Journal of Product Innovation Management*, 9 (3), (September), 171–187.

Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science*, 12, 1, (Winter), 1–27.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum (2004), "Integrating Topics and Syntax," *Advances In Neural Information Processing Systems*, 17, 537–544.

Hauser, John R. and Don Clausing (1988), "The House of Quality," *Harvard Business Review*, 66, 3, (May-June), 63–73.

Herrmann, Andreas., Frank Huber, and Christine Braunstein (2000), "Market-Driven Product and Service Design: Bridging the Gap Between Customer Needs, Quality Management, and Customer Satisfaction." *International Journal of Production Economics*, 66 (1), 77–96.

Iyyer, Mohit, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III (2015), "Deep Unordered Composition Rivals Syntactic Methods for Text Classification," *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (Volume 1: Long Papers), 1681–1691. (Beijing, China: Association for Computational Linguistics).

Kao Group (2016). http://www.company-histories.com/Kao-Corporation-Company-History.html.

Kim, Yoon (2014), "Convolutional Neural Networks for Sentence Classification," arXiv:1408.5882v2 [cs.CL], Sept 3.

Kiss, Tibor and Jan Strunk (2006), "Unsupervised Multilingual Sentence Boundary Detection," *Computational Linguistics*, 32(4), 485–525.

Le, Quoe and Tomas Mikolov (2014), "Distributed Representations of Sentences and Documents," *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 32, 1188–1196.

Lee, Thomas Y. and Eric T. Bradlow (2011), "Automated Marketing Research Using Online Customer Reviews," *Journal of Marketing Research*, 48(5), 881–894.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781v3 [cs.CL]m Sept 7,1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Che, Greg S. Corrado, and Jeffrey Dean (2013b), "Distributed Representations of Words and Phrases and Their Compositionality," *Advances In Neural Information Processing Systems*, 26, 3111–3119.

Sullivan, Lawrence P. (1986). "Quality Function Deployment," *Quality Progress*, 19(6), 39–50.

Tai, Kai Sheng, Richard Socher, and Christopher D. Manning (2015), "Improved Semantic Representations from Tree-structured Long Short-Term Memory Networks," arXiv:1503.00075v3 [cs.CL], May 30.

Timoshenko, Artem and John R. Hauser (2017), "Identifying Customer Needs from User Generated Content," (Cambridge, MA: MIT Sloan School of Management).

Ulrich Karl T. and Steven D. Eppinger (2016) *Product Design and Development, 6E*. (New York, NY: McGraw-Hill).

# WHAT A DIFFERENCE DESIGN MAKES

**KAREN BUROS**
*RADIUS GMR*[1]
**JEREMY CHRISTMAN**
*PROCTER & GAMBLE*[2]

## ABSTRACT

This case study illustrates the approach a researcher can take with Menu-Based Choice data using a study conducted by P&G. The procedures involve all aspects of the data from data cleaning to simulation, showcasing the difficulties encountered dealing with very large CPG markets.

## OVERVIEW OF THE PROBLEM

The study was undertaken to explore alternative ways P&G might bring greater organization to a large, cumbersome category. The category encompasses three compatible product types used together to accomplish a task. While this paper disguises the category under study there are many situations like this in the consumer world; laundry detergent/fabric softener/stain remover, shampoo/conditioner/hair treatment are two examples.

Three major national brands account for the clear majority of sales. The research study included only these three brands excluding smaller or regional brands. The products offered by these brands deliver specific benefits across the three product types but are not marketed as a single "bundle." A key question to be addressed is whether the consumer selects products within a "benefit space," within a brand regardless of the "benefit space" or makes differing selections from one product type to another.

A total of 86 different products were included in the study, disregarding package size, multi-packs or other promotional elements. All products were shown in comparable sizes and were priced. It should be noted that these are adult, not child-oriented, products. The products are sold in grocery, drug and mass merchandisers which often offer differing shelf sets depending on the shelf space allocated to these items.

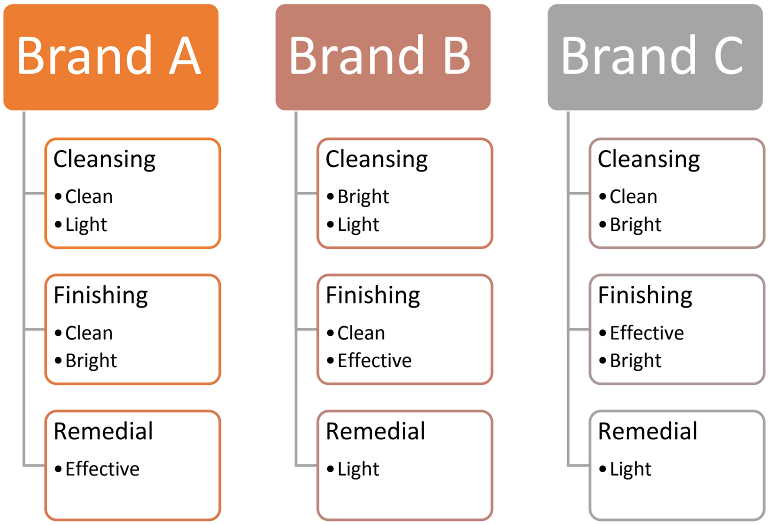
This paper will refer to the three product types as "Cleansing Products," "Finishing Products" and "Remedial Products" and the brands as A, B, C and D. Brand A offers all three product types under a common logo, Brand B utilizes a slightly modified logo in one category. Brand C offers only two product types and Brand D is a market leader in one product type.

The following table illustrates the size and complexity of the category:

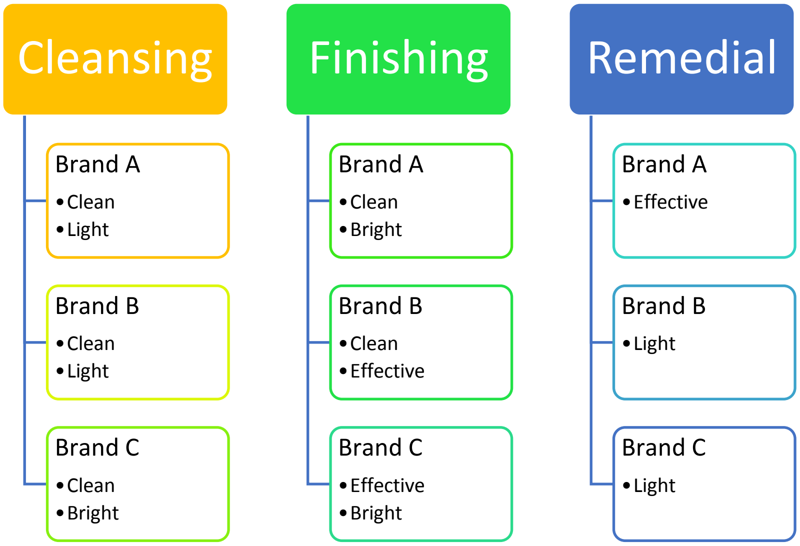
[1] Former Director Advanced Analytics Radius GMR
[2] P&G Quantitative Sciences

|  |  | Benefit Space |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | # items | Base | Clean | Clean + | Clean ++ | Bright | Effective | Multi-Function | Light | Light + | Fresh | Powerful | Mild | Mild + |
| **Cleansing** | **29** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Brand A | 15 | 5 | 5 | 0 | 0 | 1 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| Brand B1 | 12 | 1 | 6 | 0 | 0 | 2 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| Brand C | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Finishing** | **41** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Brand A | 15 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 8 | 0 | 1 | 0 | 1 | 0 |
| Brand B2 | 22 | 1 | 3 | 1 | 0 | 3 | 1 | 3 | 5 | 1 | 0 | 1 | 2 | 1 |
| Brand C | 4 | 0 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Remedial** | **16** |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Brand A | 3 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| Brand B2 | 8 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 3 | 0 | 1 | 0 | 0 | 0 |
| Brand D | 5 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Benefit Space** | **86** | **9** | **18** | **3** | **1** | **7** | **7** | **10** | **23** | **1** | **2** | **1** | **3** | **1** |

Following are two potential ways consumers might "shop" this category. In this first approach, if the consumer is shopping within brand, a retailer might place all product types of the same brand together.



Alternatively, within a benefit space, products might best be grouped by benefit.



## Research Design

The survey was conducted among 1,500+ panel members in early 2016. To qualify for the study participants were screened to be recent users of the category.

The survey used a menu-based design, meaning that each respondent saw 22 products "at price" on a screen and were asked to select the products they would likely buy from that array. Respondents were asked to select products for the household (noting that these are adult-oriented items) and could select as many or as few, including none, as they wished. They could select only one of each item.

The design was constructed such that all product types and all brands were shown in roughly equal proportions across the tasks.

| **Items within each grouping shown in roughly equal proportions across the screens** | |
|---|---|
| **Total Cleansing** | **0.24** |
| Brand A | 0.24 |
| Brand B1 | 0.24 |
| Brand C | 0.24 |
| **Total Finishing** | **0.27** |
| Brand A | 0.28 |
| Brand B2 | 0.27 |
| Brand C | 0.26 |
| **Total Remedial** | **0.25** |
| Brand A | 0.25 |
| Brand B2 | 0.25 |
| Brand D | 0.25 |

The design did not consider the overall objectives of understanding whether products were selected by benefit space or brand across product types. For example, a design could have been created to reflect groupings of items by benefit space on some screens, by brand on other screens and "random" displays on other screens thus obtaining a data-driven perspective on these issues.

## BUILDING SIMULATIONS

Given the design, HB models were built using the following criteria:

- Dependent variables are product chosen or not (86 models)
- Independent variables are product shown or not
- Cross-effects are other products shown (chi-square relationship P-value ≤.15)

In the initial simulations, all products were included "as available" so that results could be compared to known market data and raw counts from the tasks completed. While the comparison to raw counts is a bit "apples to oranges" comparison, it nonetheless provided a useful comparison in tracking down the problems in the model. The initial results were deemed nonsensical, reflecting neither market data nor "raw count" data from the tasks.

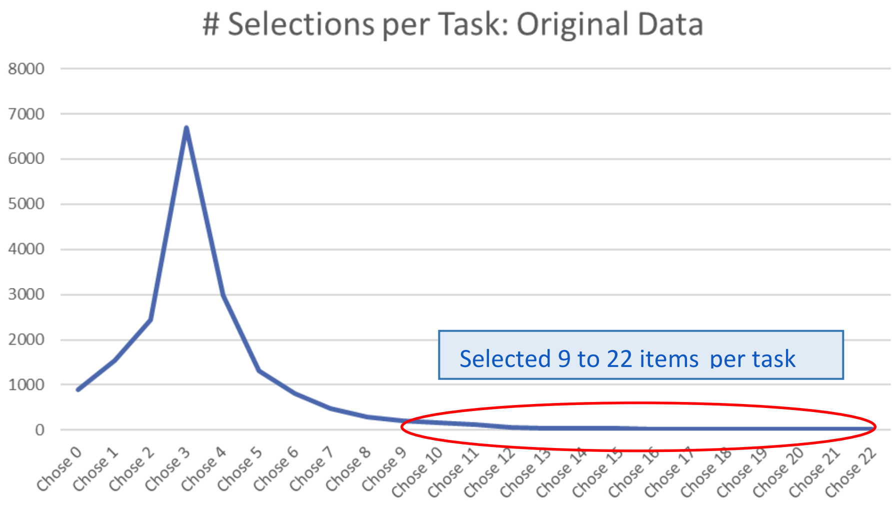
| **“Shares” from HB Simulation** | | |
|---|---|---|
| **Average # Cleansing Chosen** | **0.5** | **Too few** |
| Sum Brand A shares | 0.23 | |
| Sum Brand B1 shares | 0.16 | |
| Sum Brand C shares | 0.06 | |
| **Average # Finishing Chosen** | **0.7** | **Too few** |
| Sum Brand A shares | 0.52 | |
| sum Brand B2 shares | 0.14 | **Wrong!** |
| Sum Brand C shares | 0.00 | |
| **Average # Remedial Chosen** | **2.6** | **Too many** |
| Sum Brand A shares | 0.35 | |
| Sum Brand B2 shares | 1.79 | **Wrong!** |
| Sum Brand D shares | 0.50 | |

| **“Apples to Oranges” Point of Reference to Raw Data** | **Average # chosen/task** |
|---|---|
| Any Cleansing | 1.3 |
| Any Finishing | 1.4 |
| Any Remedial | 0.9 |

## Uncovering the Issues in the Data and the Analytic Approach

One of the goals of this paper is to illustrate the issues that should be investigated to assure a quality analysis. The first step investigates the “goodness” of the response to the survey.

Examining the responses to the individual tasks, approximately 12% of the tasks involved selection of more than eight products and were deleted from further analysis.



The next hypothesis was that models could be over-specified. Only cross-effects < .05 were included.

| Dependent variable: Item3, Chosen | | | | | |
|---|---|---|---|---|---|
| Independent Variable | Relationship P-Value | | Independent Variable | Relationship P-Value | |
| shown1 | 0.02 | | shown32 | 0.06 | x |
| shown2 | 0.02 | | shown36 | 0.03 | |
| shown3 | 0.00 | | shown37 | 0.10 | x |
| shown4 | 0.02 | | shown43 | 0.00 | |
| shown6 | 0.11 | x | shown46 | 0.14 | x |
| shown8 | 0.09 | x | shown47 | 0.14 | x |
| shown10 | 0.00 | | shown49 | 0.13 | x |
| shown11 | 0.12 | x | shown51 | 0.09 | x |
| shown14 | 0.15 | x | shown52 | 0.00 | |
| shown16 | 0.13 | x | shown55 | 0.12 | x |
| shown17 | 0.09 | x | shown62 | 0.05 | x |
| shown18 | 0.09 | x | shown67 | 0.14 | x |
| shown20 | 0.03 | | shown68 | 0.07 | x |
| shown22 | 0.03 | | shown69 | 0.05 | x |
| shown24 | 0.00 | | shown75 | 0.06 | x |
| shown25 | 0.00 | | shown77 | 0.13 | x |
| shown26 | 0.03 | | shown79 | 0.00 | |
| shown27 | 0.11 | x | shown81 | 0.06 | x |
| shown29 | 0.04 | | shown82 | 0.08 | x |
| shown31 | 0.05 | x | shown85 | 0.10 | x |

These actions resulted in only minor improvements in the simulated data when all models were re-run:

| | **HB Simulations** | | |
|---|---|---|---|
| | **Original Data** | **Cleaned/ Trimmed Data** | |
| **Average # Cleansing** | **0.45** | **0.67** | **X** |
| Sum Brand A shares | 0.23 | 0.17 | **X** |
| Sum Brand B1 shares | 0.16 | 0.44 | **X** |
| Sum Brand C shares | 0.06 | 0.06 | |
| **Average # Finishing** | **0.66** | **0.57** | **X** |
| Sum Brand A shares | 0.52 | 0.31 | |
| sum Brand B2 shares | 0.14 | 0.24 | |
| Sum Brand C shares | 0.00 | 0.02 | |
| **Average # Remedial** | **2.64** | **1.36** | **X** |
| Sum Brand A shares | 0.35 | 0.10 | **X** |
| Sum Brand B2 shares | 1.79 | 1.04 | **X** |
| Sum Brand D shares | 0.50 | 0.22 | |

Further examination of the models showed that despite the more limited number of cross-effects, the effects themselves were often counter-intuitive. To remedy the situation, logical constraints were imposed.

- **Within a category** (e.g., Cleansing), the presence of "same brand" and "other brand" items should have a negative impact on the likelihood of selection—constrain negative
- **Across categories** (e.g., Cleansing and Finishing), the presence of "same brand" items may have a positive impact—constrain positive

Once again, this resulted in some improvement but not sufficient to warrant further analysis.

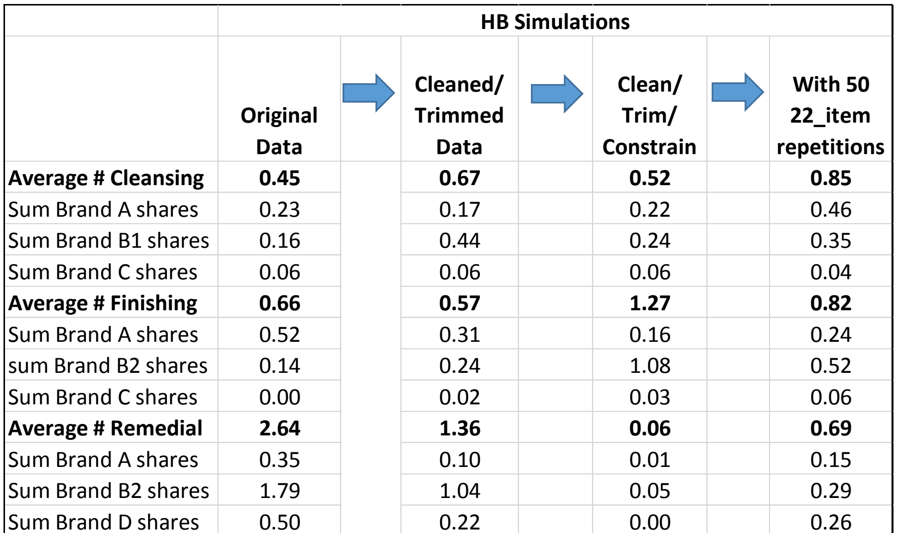
| | **HB Simulations** | | |
|---|---|---|---|
| | | **Cleaned/** | **Clean/** |
| | **Original** | **Trimmed** | **Trim/** |
| | **Data** | **Data** | **Constrain** |
| **Average # Cleansing** | **0.45** | **0.67** | **0.52** |
| Sum Brand A shares | 0.23 | 0.17 | 0.22 |
| Sum Brand B1 shares | 0.16 | 0.44 | 0.24 |
| Sum Brand C shares | 0.06 | 0.06 | 0.06 |
| **Average # Finishing** | **0.66** | **0.57** | **1.27** |
| Sum Brand A shares | 0.52 | 0.31 | 0.16 |
| sum Brand B2 shares | 0.14 | 0.24 | 1.08 |
| Sum Brand C shares | 0.00 | 0.02 | 0.03 |
| **Average # Remedial** | **2.64** | **1.36** | **0.06** |
| Sum Brand A shares | 0.35 | 0.10 | 0.01 |
| Sum Brand B2 shares | 1.79 | 1.04 | 0.05 |
| Sum Brand D shares | 0.50 | 0.22 | 0.00 |

Finally, having cleaned and constrained both the models and the data, the only place to look for the problem was in the simulations themselves. A "rule of thumb" in other choice modeling work is to make the choice task as realistic as possible to the respondent—to give as much information to the respondent such that he/she can make a realistic choice but not so much that the task becomes overwhelming.

In this very large market, the consumer has a very large array of products to choose from in a drugstore, mass merchandiser or supermarket. To keep the task from becoming overwhelming, in this case, only 22 products were shown on the screen. The extrapolation of a 22-item task to an 86-product simulation proved to be the problem. On the other hand, a 22-item simulation would not be close to a realistic simulation of the market. To explore this issue 50 tasks were selected from the original design, each showing 22 items. The selection was made by random selection of the version (out of 200 versions) and then random selection of the task (out of 12 tasks) to serve as "profiles" in simulation. The selection of profiles was checked for any bias against the original design.

| % times shown out of all tasks | Full Design | 50 Tasks |
|---|---|---|
| **Total Cleansing** | **0.24** | **0.24** |
| Brand A | 0.24 | 0.23 |
| Brand B1 | 0.24 | 0.25 |
| Brand C | 0.24 | 0.27 |
| **Total Finishing** | **0.27** | **0.27** |
| Brand A | 0.28 | 0.30 |
| Brand B2 | 0.27 | 0.25 |
| Brand C | 0.26 | 0.30 |
| **Total Remedial** | **0.25** | **0.25** |
| Brand A | 0.25 | 0.26 |
| Brand B2 | 0.25 | 0.25 |
| Brand D | 0.25 | 0.25 |

The 50 simulations were run and the average "shares" calculated across the 50 results. The difference in the results is dramatic as shown here.



|  | **HB Simulations** |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
|  | **Original Data** | → | **Cleaned/ Trimmed Data** | → | **Clean/ Trim/ Constrain** | → | **With 50 22_item repetitions** |
| **Average # Cleansing** | **0.45** |  | **0.67** |  | **0.52** |  | **0.85** |
| Sum Brand A shares | 0.23 |  | 0.17 |  | 0.22 |  | 0.46 |
| Sum Brand B1 shares | 0.16 |  | 0.44 |  | 0.24 |  | 0.35 |
| Sum Brand C shares | 0.06 |  | 0.06 |  | 0.06 |  | 0.04 |
| **Average # Finishing** | **0.66** |  | **0.57** |  | **1.27** |  | **0.82** |
| Sum Brand A shares | 0.52 |  | 0.31 |  | 0.16 |  | 0.24 |
| sum Brand B2 shares | 0.14 |  | 0.24 |  | 1.08 |  | 0.52 |
| Sum Brand C shares | 0.00 |  | 0.02 |  | 0.03 |  | 0.06 |
| **Average # Remedial** | **2.64** |  | **1.36** |  | **0.06** |  | **0.69** |
| Sum Brand A shares | 0.35 |  | 0.10 |  | 0.01 |  | 0.15 |
| Sum Brand B2 shares | 1.79 |  | 1.04 |  | 0.05 |  | 0.29 |
| Sum Brand D shares | 0.50 |  | 0.22 |  | 0.00 |  | 0.26 |

As a final step, the exponent in the simulator is adjusted to fine-tune the results to more closely reflect known market data and the data from the raw counts.

| Counts from Tasks Cleaned Data | |
|---|---|
| **Total Cleansing** | **1.13** |
| Brand A | 0.62 |
| Brand B1 | 0.47 |
| Brand C | 0.04 |
| **Total Finishing** | **1.18** |
| Brand A | 0.32 |
| Brand B2 | 0.79 |
| Brand C | 0.08 |
| **Total Remedial** | **0.79** |
| Brand A | 0.15 |
| Brand B2 | 0.38 |
| Brand D | 0.26 |

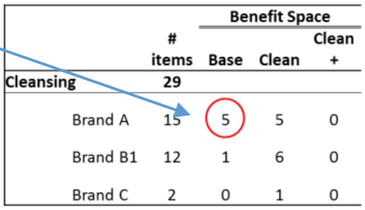
| | **HB Simulations** | |
|---|---|---|
| | **50 22_item repetitions** | |
| **Tuning Exponent** | **1.0** | **0.8** |
| **Average # Cleansing** | **0.82** | **1.00** |
| Sum Brand A shares | 0.43 | 0.54 |
| Sum Brand B1 shares | 0.35 | 0.43 |
| Sum Brand C shares | 0.03 | 0.03 |
| **Average # Finishing** | **0.86** | **1.18** |
| Sum Brand A shares | 0.24 | 0.32 |
| sum Brand B2 shares | 0.57 | 0.78 |
| Sum Brand C shares | 0.05 | 0.08 |
| **Average # Remedial** | **0.71** | **0.76** |
| Sum Brand A shares | 0.14 | 0.15 |
| Sum Brand B2 shares | 0.34 | 0.36 |
| Sum Brand D shares | 0.23 | 0.24 |

## RESOLVING THE DESIGN ISSUES

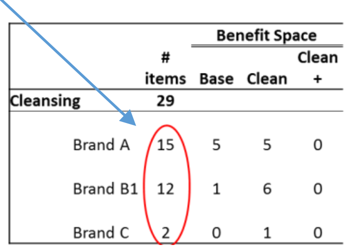
Given an understanding of the data and simulation issues, the question becomes whether this design could be better utilized to answer the original questions of selection of product by brand or benefit space.

To explore this, the selections made by the respondent could be recoded in several ways. To illustrate the concept consider the following:

- Original Coding: Dummy-coded: Chosen/Not Chosen
- Combinatorial: groups of items chosen together (e.g., peanut butter and jelly)
- Brand_Benefit within Category
- Brand within category



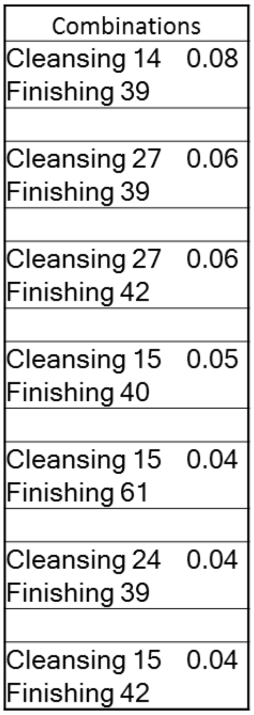
| | | Benefit Space | | |
|---|---|---|---|---|
| | # items | Base | Clean | Clean + |
| **Cleansing** | **29** | | | |
| Brand A | 15 | 5 | 5 | 0 |
| Brand B1 | 12 | 1 | 6 | 0 |
| Brand C | 2 | 0 | 1 | 0 |



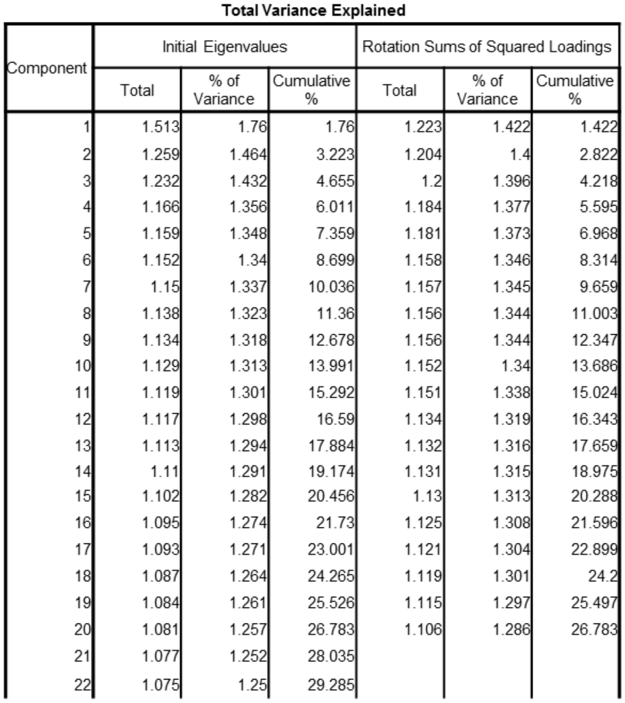
| | | Benefit Space | | |
|---|---|---|---|---|
| | # items | Base | Clean | Clean + |
| **Cleansing** | **29** | | | |
| Brand A | 15 | 5 | 5 | 0 |
| Brand B1 | 12 | 1 | 6 | 0 |
| Brand C | 2 | 0 | 1 | 0 |

For the combinatorial approach, three alternatives are explored. First, a straight count of the number of times items are chosen together in a task is made. It is not surprising that no combinations of items appear to dominate given the random selection of 22 items shown in a task.

Then a principal components factor analysis was attempted, again with no meaningful combinations emerging from the data.



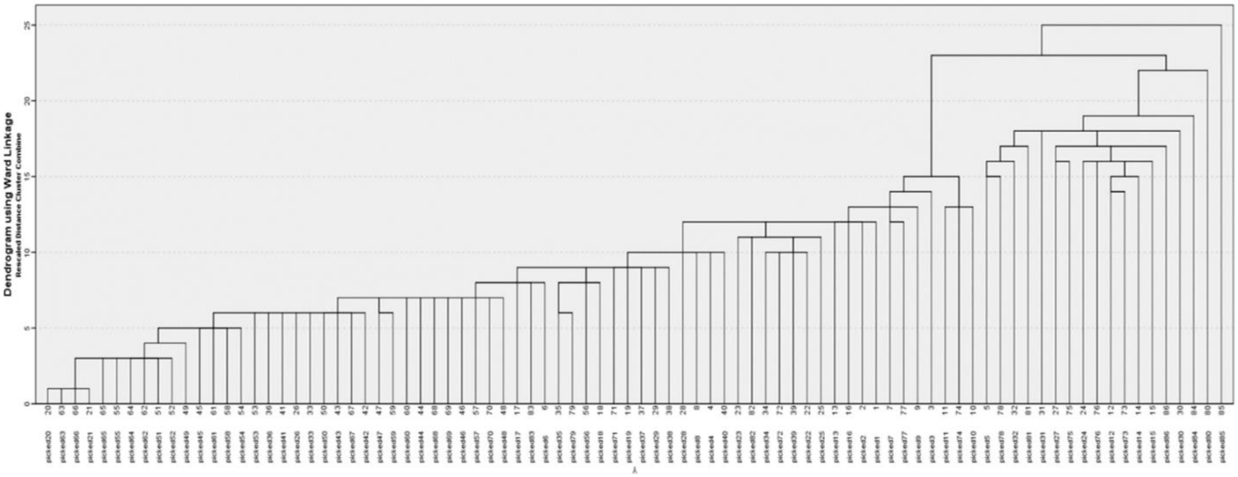
| Combinations |
|---|
| Cleansing 14 0.08<br>Finishing 39 |
| |
| Cleansing 27 0.06<br>Finishing 39 |
| |
| Cleansing 27 0.06<br>Finishing 42 |
| |
| Cleansing 15 0.05<br>Finishing 40 |
| |
| Cleansing 15 0.04<br>Finishing 61 |
| |
| Cleansing 24 0.04<br>Finishing 39 |
| |
| Cleansing 15 0.04<br>Finishing 42 |



**Total Variance Explained**

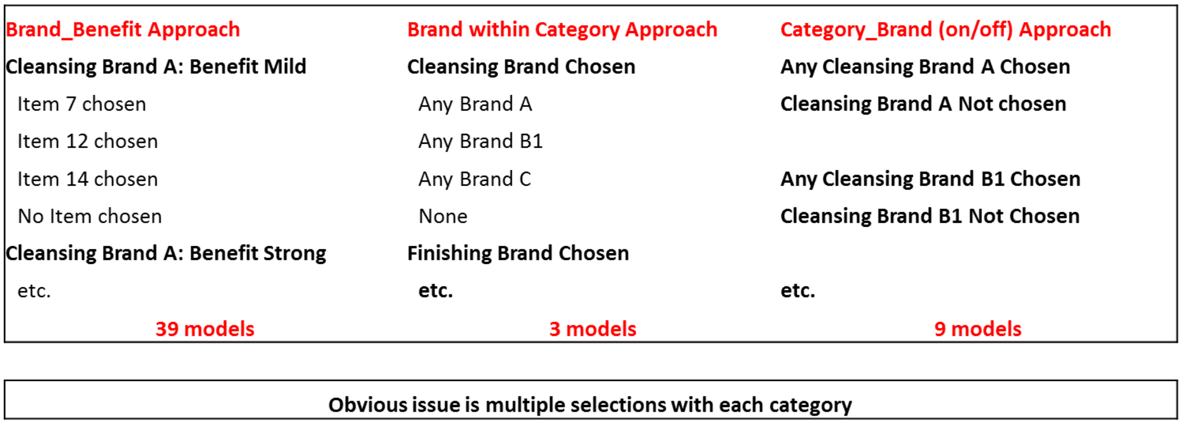
| Component | Initial Eigenvalues | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 1.513 | 1.76 | 1.76 | 1.223 | 1.422 | 1.422 |
| 2 | 1.259 | 1.464 | 3.223 | 1.204 | 1.4 | 2.822 |
| 3 | 1.232 | 1.432 | 4.655 | 1.2 | 1.396 | 4.218 |
| 4 | 1.166 | 1.356 | 6.011 | 1.184 | 1.377 | 5.595 |
| 5 | 1.159 | 1.348 | 7.359 | 1.181 | 1.373 | 6.968 |
| 6 | 1.152 | 1.34 | 8.699 | 1.158 | 1.346 | 8.314 |
| 7 | 1.15 | 1.337 | 10.036 | 1.157 | 1.345 | 9.659 |
| 8 | 1.138 | 1.323 | 11.36 | 1.156 | 1.344 | 11.003 |
| 9 | 1.134 | 1.318 | 12.678 | 1.156 | 1.344 | 12.347 |
| 10 | 1.129 | 1.313 | 13.991 | 1.152 | 1.34 | 13.686 |
| 11 | 1.119 | 1.301 | 15.292 | 1.151 | 1.338 | 15.024 |
| 12 | 1.117 | 1.298 | 16.59 | 1.134 | 1.319 | 16.343 |
| 13 | 1.113 | 1.294 | 17.884 | 1.132 | 1.316 | 17.659 |
| 14 | 1.11 | 1.291 | 19.174 | 1.131 | 1.315 | 18.975 |
| 15 | 1.102 | 1.282 | 20.456 | 1.13 | 1.313 | 20.288 |
| 16 | 1.095 | 1.274 | 21.73 | 1.125 | 1.308 | 21.596 |
| 17 | 1.093 | 1.271 | 23.001 | 1.121 | 1.304 | 22.899 |
| 18 | 1.087 | 1.264 | 24.265 | 1.119 | 1.301 | 24.2 |
| 19 | 1.084 | 1.261 | 25.526 | 1.115 | 1.297 | 25.497 |
| 20 | 1.081 | 1.257 | 26.783 | 1.106 | 1.286 | 26.783 |
| 21 | 1.077 | 1.252 | 28.035 | | | |
| 22 | 1.075 | 1.25 | 29.285 | | | |

Finally, a hierarchical clustering of the items was tried. Again, no meaningful clusters emerge.



Recoding the choices made by the respondent into a simplified design space appear to be the only available alternatives. The obvious disadvantage of this approach is that the respondent could select more than one of the same type item falling within a “recoded” space, e.g., brand. That said, it might provide some understanding of the selection process in the analysis stage. The reduced number of models could possibly yield greater stability in simulation.

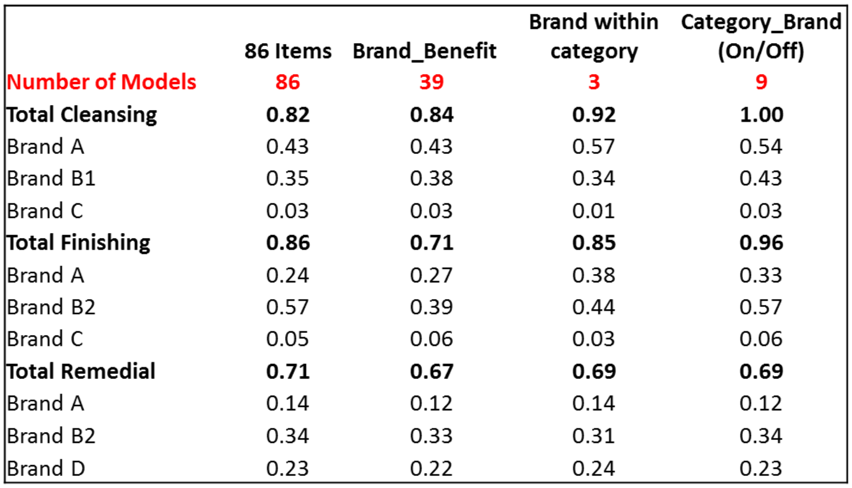
Three approaches were undertaken as follows:



| Brand_Benefit Approach | Brand within Category Approach | Category_Brand (on/off) Approach |
|---|---|---|
| **Cleansing Brand A: Benefit Mild** | **Cleansing Brand Chosen** | **Any Cleansing Brand A Chosen** |
| Item 7 chosen | Any Brand A | **Cleansing Brand A Not chosen** |
| Item 12 chosen | Any Brand B1 | |
| Item 14 chosen | Any Brand C | **Any Cleansing Brand B1 Chosen** |
| No Item chosen | None | **Cleansing Brand B1 Not Chosen** |
| **Cleansing Brand A: Benefit Strong** | **Finishing Brand Chosen** | |
| etc. | **etc.** | **etc.** |
| 39 models | 3 models | 9 models |

**Obvious issue is multiple selections with each category**

Again, using the simulation averaging approach (50 simulations of 22 items) the results of these recoded models can be compared to the 86-item simulation. All approaches lead to very similar results. The choice of which approach to use becomes purely an analytic choice, noting that none of these completely resolves the issue of the random nature of the original design.



| | 86 Items | Brand_Benefit | Brand within category | Category_Brand (On/Off) |
|---|---|---|---|---|
| Number of Models | 86 | 39 | 3 | 9 |
| **Total Cleansing** | **0.82** | **0.84** | **0.92** | **1.00** |
| Brand A | 0.43 | 0.43 | 0.57 | 0.54 |
| Brand B1 | 0.35 | 0.38 | 0.34 | 0.43 |
| Brand C | 0.03 | 0.03 | 0.01 | 0.03 |
| **Total Finishing** | **0.86** | **0.71** | **0.85** | **0.96** |
| Brand A | 0.24 | 0.27 | 0.38 | 0.33 |
| Brand B2 | 0.57 | 0.39 | 0.44 | 0.57 |
| Brand C | 0.05 | 0.06 | 0.03 | 0.06 |
| **Total Remedial** | **0.71** | **0.67** | **0.69** | **0.69** |
| Brand A | 0.14 | 0.12 | 0.14 | 0.12 |
| Brand B2 | 0.34 | 0.33 | 0.31 | 0.34 |
| Brand D | 0.23 | 0.22 | 0.24 | 0.23 |

The initial hypothesis was that the 86-item model was the source of the problematic results and that simplification of the models would resolve the issue. This proved not to be the case. When the alternative coded models were simulated using all 86 items, not only were the results contrary to the ingoing data and known market dynamics but they also differed markedly from one to the other. Once the data issues and, most importantly, the simulation issues were resolved the models fell into line as one would expect.

It is possible that respondent-level clustering may reveal underlying patterns of choice. That approach was not explored in this case study. An alternative design, focusing on brands and benefit spaces in the tasks would more likely result in more meaningful findings.

## Learnings

As painful as this exercise proved to be, it nonetheless yields some meaningful learnings when approaching menu-based choice modeling.

- Sense check the data and trim out the poor responses
- Conceptualize the models to mirror the objectives of the analysis
- Use the tools you have to obtain the best model . . .
  - Minimize cross-effects
  - Constrain cross-effects when it makes sense
  - Calibrate—fine-tune—your exponent
- Bring the simulator in line with the respondent task—or vice versa
- Give yourself enough time to do the job right

Karen Buros

Jeremy Christman

# EXPLAINING PREFERENCE HETEROGENEITY WITH MIXED MEMBERSHIP MODELING

**MARC R. DOTSON**
*BRIGHAM YOUNG UNIVERSITY*
**JOACHIM BÜSCHKEN**
*CATHOLIC UNIVERSITY OF EICHSTÄTT-INGOLSTADT*
**GREG M. ALLENBY**
*OHIO STATE UNIVERSITY*

## 1 INTRODUCTION

The fact that consumers are heterogeneous in their preferences gives rise to marketing as a discipline and an industry. Choice models and associated decision tools that account for this heterogeneity allow firms to better understand what consumers prefer and have become a standard for product development and product line optimization. However, explaining preference heterogeneity remains an elusive problem. In this paper we develop an expanded choice model that improves our ability to explain preference heterogeneity by employing a novel approach to model discrete data, including binary and ratings survey data, that describe the drivers of consumer preference.

Choice modeling is an effective tool for determining what product attributes individuals prefer but it has proven less successful at explaining the heterogeneity in consumer preferences. Explaining preference heterogeneity includes identifying covariates that serve as drivers of preference and enable targeting and promotion activities. The use of hierarchical Bayes in choice modeling allows for both individual-level attribute part-worth utilities and aggregate-level preference heterogeneity parameters. Part-worth estimates tell us what attributes consumers prefer. Parameters describing preference heterogeneity are conditioned on covariates that help explain cross-sectional variation in the part-worths.

Finding covariates that are predictive of part-worths has proven difficult. The primary benefit when using a random effect distribution of heterogeneity has been accounting for unexplained heterogeneity. Using discrete variables describing possible drivers of preference, such as demographics and psychographics, as covariates is standard. However, survey data are typically used as covariates where the number of covariates makes it impractical to include interactions. Additionally, we have growing access to new sources of discrete multivariate data outside of surveys, including text, that we expect will be a rich source of information for explaining choice yet incorporating it isn’t obvious. We propose modeling this discrete multivariate data as part of the choice model in order to uncover covariates that can better explain preference heterogeneity.

In this paper we develop an expanded hierarchical Bayesian choice model where covariates for the upper level are from a grade of membership model (Woodbury *et al.* 1978, Erosheva *et al.* 2007). The grade of membership model is related to latent Dirichlet allocation, which serves as a touchstone within topic modeling (Blei *et al.* 2003). Both are part of a larger class of models known as mixed membership models that provide individual-level, low-dimensional representations of discrete multivariate data by accounting for interactions or co-occurrence (Airoldi *et al.* 2014). We propose modeling discrete variables describing potential drivers of preference where interaction among drivers will help further explain preference heterogeneity. We apply our model within the robotic vacuums category and find we can both explain preference heterogeneity and predict choice better than traditional models using observed covariates directly.

This paper contributes to efforts at using mixed membership models to improve marketing models. The application of this class of models to marketing contexts is still in its infancy. Extant research has focused on latent Dirichlet allocation (LDA), using product reviews and online forums to inform market structure (Lee and Bradlow 2011, Netzer *et al.* 2012) and to identify preferences for product features (Archak *et al.* 2011). Most recently, Tirunillai and Tellis (2014) use LDA to conduct brand analysis while Büschken and Allenby (2016) develop a sentence-constrained LDA to better predict review ratings. However, mixed membership models have yet to be employed in the context of choice modeling. We believe this paper provides an important first step in this regard.

The remainder of the paper will be organized as follows. We specify our model in Section 2. We detail our empirical application in Section 3. In Section 4, we compare results from our proposed model, with covariates uncovered using the grade of membership model, and alternative models where standard discrete covariates are used. We discuss implications of and extensions to this research in Section 5.

## 2 Model Specification

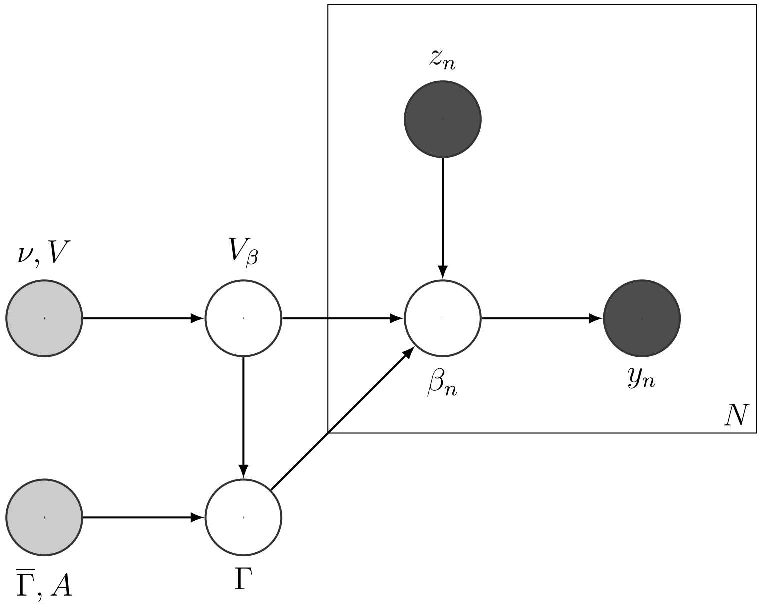
### 2.1 Hierarchical Bayesian Choice Model

Hierarchical Bayesian choice models allow for the estimation of both individual and aggregate-level preference parameters, even in the presence of few observations per individual (Rossi and Allenby 2003, Rossi *et al.* 2005). Decision tools associated with choice modeling make use of individual-level preference parameter estimates to forecast the results of various product policies while aggregate-level parameter estimates are employed to explain the source of individual preferences.

The likelihood in hierarchical Bayesian choice modeling is typically assumed to be a multinomial logit model such that the probability of an individual choosing a product alternative is a function of the attributes that compose the given alternative and the part-worths or individual-level preferences for the attributes. The distribution of heterogeneity, or upper level, models preference heterogeneity in the individual-level part-worths. The distribution of heterogeneity is typically assumed to be multivariate normal. The mean of the distribution of heterogeneity is where the analyst can specify individual-specific covariates that explain variation in the part-worths.

The directed acyclic graph (DAG) in Figure 1 provides a visual representation of the hierarchical Bayesian choice model. The DAG utilizes plate notation, where a plate represents replication for the enclosed variables. In the DAG, white nodes represent parameters to be estimated, grey nodes represent fixed hyper-parameters, and black nodes represent observed data. From the use of plate notation in Figure 1, we can see that the hierarchical Bayesian choice model has both aggregate and individual levels. To be clear, at the aggregate level, we have the hyper-parameters for conjugate normal ($\Gamma$-bar and $A$) and inverse Wishart ($\nu$ and $V$) priors and the parameters for the distribution of heterogeneity ($V_\beta$ and $\Gamma$). At the individual level, $y_n$ is a vector of observed choices, $\beta_n$ are the part-worths, and $z_n$ are the observed covariates for individual $n$. We can see that the covariates are chosen independent of the model specification. As discussed, the covariates are the key to our ability to explain preference heterogeneity. We will use DAGs, beginning with Figure 1, to help motivate the proposed model.

**Figure 1. Hierarchical Bayesian Choice Model**



A variety of covariates have been employed to explain preference heterogeneity in the choice modeling literature. For example, Allenby and Ginter (1995) used demographic variables, Lenk *et al.* (1996) included expertise, and Chandukala *et al.* (2011) specified consumer needs to explain variation in the part-worths. However, explaining preference heterogeneity has not met with much success generally (Rossi *et al.* 1996, Horsky *et al.* 2006).

One unresolved issue is that discrete covariates are often employed without a practical way to include interactions. The problem is one of dimensionality. The number of interaction terms is $J$ choose $M$, where $J$ is the number of covariates and $M$ is the number of desired interactions. For example, with $J = 30$ covariates and $M = 2$, there are 435 possible two-way interactions, to say nothing of higher-level interactions where $M > 2$. While Chandukala *et al.* (2011) employ variable selection to determine which covariates matter, we are interested in a model general enough to account for interactions from traditional survey data as well as accommodate new sources of discrete data.

We propose using a non-standard model that accounts for the interaction or co-occurrence of variables to uncover covariates from discrete multivariate data for use in a choice model's random effect distribution of heterogeneity. Specifically, we propose combining a hierarchical Bayesian choice model with a grade of membership model to uncover covariates that account for interactions in order to explain preference heterogeneity better than using observed covariates directly. We first detail the grade of membership and the class of mixed membership models before specifying our expanded choice model.

## 2.2 The Grade of Membership Model

The grade of membership (GoM) model was developed to classify disease patterns using discrete patient-level clinical data (Woodbury *et al.* 1978, Clive *et al.* 1983). It has since been applied to modeling survey data (Erosheva *et al.* 2007, Gross and Manrique-Vallier 2014). In these applications, each respondent answers a battery of survey questions with categorical responses. The research interest is to identify the patterns of interaction or co-occurrence in the categorical responses across respondents along with how each respondent relates to the patterns of co-occurrence. The GoM model characterizes these patterns of co-occurrence as profiles of archetypal respondents. Each respondent is a partial member of each of the profiles based on how similar their responses are to each pattern of co-occurrence.

**Figure 2. Modeling Pick Any/J Data with a GoM Model**

(a) Respondent $n$'s Responses and Membership Vector $g_n$

| *What benefits does cereal provide that are important to you?* |
|---|
| Item 1: It's a helpful way to get a serving of milk at the same time<br>Item 2: Cereal is a good source of fiber<br>Item 3: My kids will eat cereal for breakfast<br>Item 4: Cereal isn't just for breakfast, it's a good snack anytime<br>Item 11: I want to make sure my family has breakfast in the morning<br>Item 15: Cereal is easy to prepare |
| $g_n$ "Kids Breakfast" 0.60 "Healthy Snack" 0.20 "Source of Fiber" 0.20 |

(b) Aggregate-Level Profiles Defined by the Probability of Each Item $\lambda_{j,k}(1)$

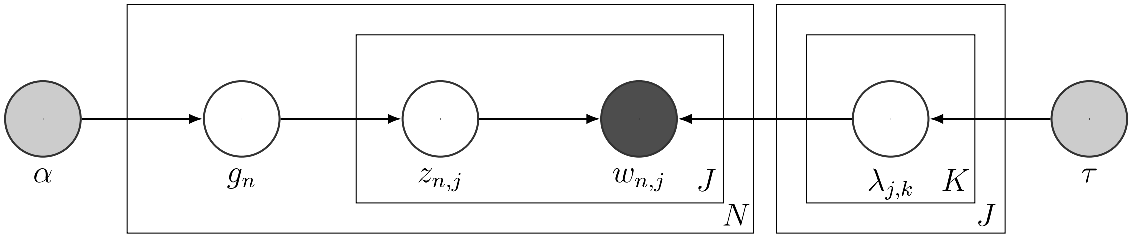
| $\lambda_{j,k}(1)$ | "Kids Breakfast" | "Healthy Snack" | "Source of Fiber" |
|---|---|---|---|
| Item 1 | 0.67 | 0.70 | 0.34 |
| Item 2 | 0.22 | 0.85 | **0.95** |
| Item 3 | **0.97** | 0.13 | 0.04 |
| Item 4 | 0.32 | **0.92** | 0.10 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| Item 30 | 0.04 | 0.13 | 0.14 |

To illustrate, consider responses to a battery of select-all-that-apply questions (i.e., pick any/*J*). Each respondent selects or indicates a subset of the $J = 30$ statements or items that apply to them in answer to the question: "What benefits does cereal provide that are important to you?" Figure 2(a) displays the items selected for a given respondent together with their membership vector $g_n$. Figure 2(b) displays $\lambda_{j,k}(1)$ describing $K = 3$ aggregate-level profiles in terms of the likelihood of selecting each of the $J = 30$ items. Note that since there are only two categorical options for all $J = 30$ questions, each $\lambda_{j,k}$ is a vector with two elements such that $\lambda_{j,k}$ (0) is the complement of the values listed in Figure 2(b). Thus $\lambda_{j,k}(0) + \lambda_{j,k}(1) = 1$ for each $\lambda_{j,k}$.

Using Figure 2, we can see how profiles emerge based on what items co-occur. For example, if item 11 "I want to make sure my family has breakfast in the morning" and item 3 "My kids will eat cereal for breakfast" are selected together frequently across respondents, this pattern may be part of a profile describing concern with breakfast for children. In Figure 2(a), the membership vector $g_n$ describes the partial membership respondent $n$ has in each of the $K = 3$ profiles—"Kids Breakfast," "Healthy Snack," and "Source of Fiber"—where the number of profiles $K = 3$ has been specified by the analyst and the weight given to each profile is determined by how similar respondent $n$'s response pattern matches each of the aggregate-level profiles. For this particular respondent, they are primarily a member of the "Kids Breakfast" profile, with a weight of 0.60, while still being a partial member of the remaining two profiles. The membership vector $g_n$ has non-negative elements and is constrained to equal 1.

The aggregate-level values $\lambda_{j,k}(1)$ in Figure 2(b) describe how likely it is for each item to occur within each profile. The profiles are composed of all $J = 30$ items with the item that is most likely within each profile in bold. Based on common response patterns across respondents, the profiles describe archetypal or extreme respondents (i.e., respondents that define the bounds of the convex hull), ones that in this case are either concerned wholly with cereal for "Kids Breakfast," a "Healthy Snack," or a "Source of Fiber," where the profile names have been determined by the analyst based on which items differentiate each profile. Thus each membership vector $g_n$ describes where a respondent $n$ is located within a convex hull defined by the extreme respondent profiles. These profiles account for the co-occurrence or interaction of the discrete items while reducing the dimensionality from $J$ to $K$.

**Figure 3. The Grade of Membership Model**



The DAG in Figure 3 provides a visual representation of the GoM model. The plate notation demonstrates the three model levels: item, respondent, and aggregate. The aggregate-level $\lambda_{j,k}$ describing profiles is homogeneous while the respondent-level membership vectors $g_n$ are heterogeneous. To be clear, the hyper-parameters are for conjugate Dirichlet priors ($\alpha$ and $\tau$) and the latent variables $z_{n,j}$ are different from the observed covariates in Figure 1.

In the marketing literature, it has been argued that identifying extreme responses is important for designing and promoting successful new products (Allenby and Ginter 1995). For example, extreme response behavior can be used to more efficiently target prospects with a high probability of adopting an innovation. Conceptualizing consumer heterogeneity as a continuous distribution of preferences has been shown to aid in the identification of extreme responses (Allenby *et al.* 1998, Allenby and Rossi 1998). The GoM model represents discrete response behavior as a continuous proximity to a limited number of extreme profiles. Given that marketers often search for a limited number of product offerings for reasons of efficiency or resource limitations, a concept of heterogeneity that expresses differences among consumers in the space of a small number of extreme response profiles is appealing. We utilize the GoM model given this characterization of heterogeneity, which includes the respondent-level membership vectors $g_n$, in the development of our proposed model.

### 2.2.1 Relationship with Finite Mixture Models

Having a respondent-level membership vector that consists of non-negative, real-valued latent variables that sum to one is the distinctive feature of mixed membership models, the class of models that includes the GoM and LDA. Contrast this with the general form of a finite mixture model (Kamakura and Russell 1989) where we have a membership vector at the aggregate level while the GoM model in has a membership vector at the individual level. This feature is common to all mixed membership models and illustrates why they are often referred to as individual-level mixture models.

Finite mixture models are a special case of mixed membership models (Erosheva *et al.* 2007, Galyardt 2014). However, our use of the GoM within the class of mixed membership models is different than the typical use of finite mixture models in choice modeling. Instead of specifying a mixture of distributions of heterogeneity, we are interested in using the respondent-level membership vector $g_n$ to serve as covariates that can further explain preference heterogeneity.

### 2.2.2 Relationship with Factor Analysis

Factor analysis is another related model and has long been a standard approach in marketing for dimension reduction (Stewart 1981). The basic assumption is that a set of variables can be reduced to one or more latent constructs called factors. The form of factor analysis is similar to that of the GoM model, with factor scores in place of the membership vector and factor loadings in place of the profiles. Erosheva (2002) even demonstrates that the GoM model is equivalent to a binary factor analysis with an identity link function. However, there are key differences in the two approaches.

Factor analysis and GoM models differ in terms of their underlying assumptions, modeling objectives, and the type of data each method can process (Manton *et al.*, 1994; Marini *et al.* 1996). First, standard factor analysis assumes continuous data. Even using a cut-point model, which assumes the observed data are discrete indicators of latent continuous variables, the underlying constructs (i.e., factors) are still considered to be continuous. On the other hand, the GoM model assumes both discrete data and discrete underlying constructs (i.e., profiles).

Second, the objective of factor analysis is to uncover latent constructs underlying a set of variables. The objective of the GoM model is to both uncover profiles representing extreme characterizations of respondents and measure each respondent’s proximity to these profiles. In other words, the GoM model has the description of respondents and respondent heterogeneity as the objects of inference. Finally, unlike factor analysis, the GoM model can handle a combination of multinomial, ordinal, and other discrete multivariate data.

## 2.3 Hierarchical Bayesian Choice Model with a GoM Model

The proposed model combines a hierarchical Bayesian choice model with a GoM model in order to use discrete multivariate data to uncover covariates that explain preference heterogeneity. A related concept is presented in the form of a supervised latent Dirichlet allocation (sLDA). In the sLDA topic model, each collection of discrete data (i.e., document, in the context of topic modeling) is paired with and used to be predictive of a response, such as using movie reviews to predict movie ratings (Blei and McAuliffe 2007). We employ the same kind of pairing between a collection of discrete data and response, however our response is part-worth utility parameters and the collection of discrete data is from a battery of survey questions.

The individual-level choice model remains multinomial logit and the distribution of heterogeneity remains multivariate normal. Since there is a separate $g_n$ for each respondent in the GoM model, we use these membership vectors as covariates to explain heterogeneity in the part-worths. Figure 4 illustrates the proposed hierarchical Bayesian choice model with a GoM model. From the DAG we can see that the proposed model is a three-level model where only the categorical responses and choices for each respondent are observed. The homogeneous profiles $\lambda$ account for the interaction or co-occurrence among items and provide for the dimension reduction we need to use this collection of discrete data as covariates in the model of preference heterogeneity.

**Figure 4. Hierarchical Bayesian Choice Model with a GoM Model**

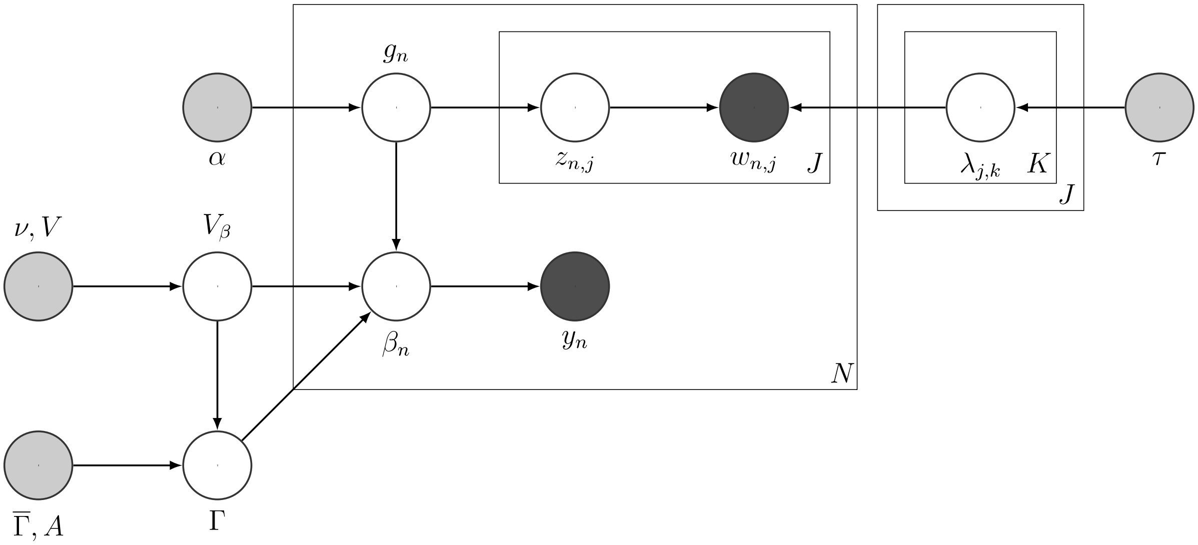


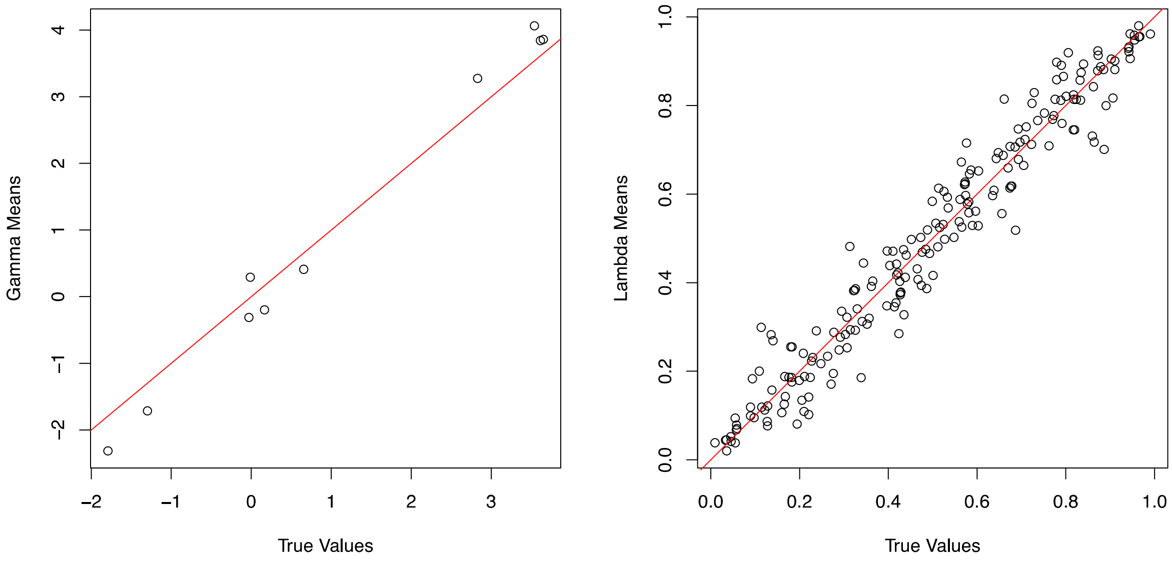
Figure 4 combines the DAGs in Figure 1 and Figure 3 to illustrate that the membership vector $g_n$ serves as the link between the choice model and the GoM model. Thus $g_n$ is informed by both the categorical responses and the chosen alternatives. The proposed model is more complete than a model where $g_n$ is estimated separately from choice since estimating all the parameters in the expanded model allows us to properly account for the uncertainty in $g_n$. A complete list of the variables in Figure 4 are detailed in Table 1.

**Table 1. Variables in Hierarchical Bayesian Choice Model with a GoM Model**

| Choice Variables | Description |
|---|---|
| $N$ | number of respondents |
| $H$ | number of choice tasks for each respondent $n$ |
| $P$ | number of alternatives in each choice task |
| $L$ | number of attribute levels in each choice task |
| $y_n$ | $H$-dim vector of choices for respondent $n$ |
| $\beta_n$ | $L$-dim vector of part-worths for respondent $n$ |
| $\Gamma$ | $K \times L$ matrix representing the mean of the random effects distribution of heterogeneity |
| $V_\beta$ | $L \times L$ covariance matrix of the random effects distribution of heterogeneity |
| GoM Variables | Description |
| $K$ | number of profiles |
| $J$ | number of categorical questions |
| $n_j$ | number of categorical responses for question $j$ |
| $w_n$ | $J$-dim vector of respondent $n$'s categorical responses |
| $z_n$ | $J$-dim vector of respondent $n$'s profile assignments |
| $g_n$ | $K$-dim membership vector for respondent $n$ |
| $\lambda$ | collection of probability distributions $\lambda_{j,k}$ over the $n_j$ response options for each question $j$ and profile $k$ |

We validated our proposed model by generating data where $K$ = 2, $N$ = 200, $J$ = 13, $n_j$ = 2 for all $J$, $H$ = 50, $P$ = 4, and $L$ = 5 and recovering parameter values. Each true parameter value was within or near the bounds of a 95% credible interval. We display the aggregate-level posterior means in Figure 5. The posterior means line up along the diagonal, indicating parameter recovery. Note that the $\lambda$ estimates are constrained to be within the 0–1 bounds.

**Figure 5. Simulation Study Results**



## 3 Empirical Application

We use data from a national survey of preferences regarding robotic vacuums. A total of 332 respondents were carefully screened to ensure that the product options under consideration were relevant to them. In particular, qualified respondents had to own a robotic vacuum, currently be shopping for their first robotic vacuum, or might consider a robotic vacuum sometime in the next five years.

Before the conjoint experiment, respondents were asked to detail why the product was relevant to them or anyone in their household by selecting from a list of 11 statements on cleaning that robotic vacuums might help address. Respondents were also asked to select from among a list of 7 statements that described problems with robotic vacuums. The combined list of 18 statements regarding cleaning and robotic vacuums is provided in Table 2. Thus our discrete data consists of two possible categories where not selecting an item is coded as a 0 and selecting an item is coded as a 1.

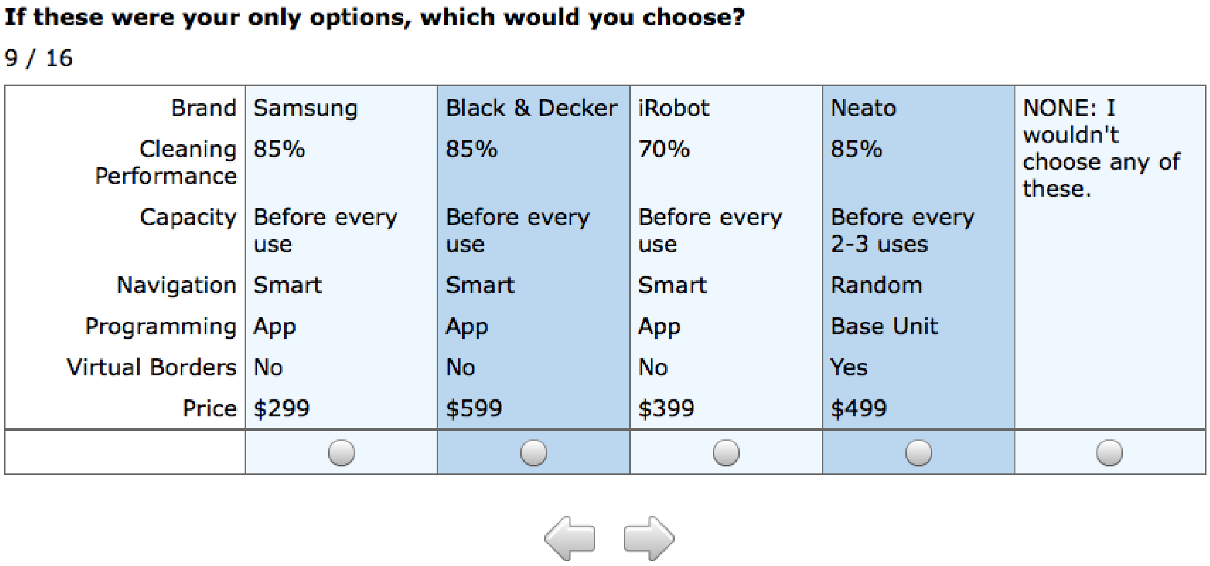
**Table 2. Statements on Cleaning and Robotic Vacuums**

| No. | Item |
|---|---|
| 1 | I enjoy coming home to a clean house. |
| 2 | I don't feel relaxed when I know my home isn't clean. |
| 3 | I worry about pet hair and dander in the home. |
| 4 | I have trouble keeping the floor beneath my furniture clean. |
| 5 | I worry about germs and dirt on my floor and carpet. |
| 6 | I get anxious about having guests when my home is dirty. |
| 7 | I don't like going to someone's home that is dirty. |
| 8 | I don't like touching dirty things. |
| 9 | I don't spend much time cleaning. |
| 10 | I spend over two hours per week cleaning. |
| 11 | I have a cleaning person who cleans for me. |
| 12 | Robotic vacuums are too expensive. |
| 13 | Robotic vacuums are too complicated to program, set up, and operate. |
| 14 | Robotic vacuums often need to be "rescued" because they get stuck. |
| 15 | Robotic vacuums need to have their trash containers changed too often. |
| 16 | Robotic vacuums don't do a good enough job cleaning the floor and carpet. |
| 17 | Robotic vacuums don't spend enough time on really dirty spots on the floor. |
| 18 | Robotic vacuums scare household pets. |

Standard models using this discrete data as observed covariates in the random effects distribution of heterogeneity don't have a practical way to include interactions, even though interactions should be expected. For example, we would expect that respondents who select statement 5 "I worry about germs and dirt on my floor and carpet" also select statement 10 "I spend over two hours per week cleaning" and that this interaction would have an impact on explaining preferences in the random effects distribution of heterogeneity. However, if we were to include two-way interactions, we would add an additional 153 covariates, to say nothing of the dimensionality introduced by higher-level interactions.

After selecting from applicable statements on cleaning and robotic vacuums, respondents proceeded through a series of 16 choice tasks where they were asked to select which of five product alternatives they most preferred, including an outside option to not select any of the given alternatives. Figure 6 is a screenshot of one of these choice tasks. Each alternative was composed of seven separate attributes for a total of 12 estimable attribute levels, excluding the reference levels in bold detailed in Table 3.

## Figure 6. Example Choice Task



**If these were your only options, which would you choose?**

9 / 16

| | | | | | |
|---|---|---|---|---|---|
| Brand | Samsung | Black & Decker | iRobot | Neato | NONE: I wouldn't choose any of these. |
| Cleaning Performance | 85% | 85% | 70% | 85% | |
| Capacity | Before every use | Before every use | Before every use | Before every 2-3 uses | |
| Navigation | Smart | Smart | Smart | Random | |
| Programming | App | App | App | Base Unit | |
| Virtual Borders | No | No | No | Yes | |
| Price | $299 | $599 | $399 | $499 | |
| | ○ | ○ | ○ | ○ | ○ |

## Table 3. Attribute Levels

| Attributes | Levels | | | | |
|---|---|---|---|---|---|
| Brand | **Outside Option** | Neato | iRobot | Samsung | Black & Decker |
| Performance | **70%** | 85% | | | |
| Capacity | **Every use** | Every 2-3 uses | | | |
| Navigation | **Random** | Smart | | | |
| Programming | **Base unit** | App | | | |
| Virtual Borders | **No** | Yes | | | |
| Price | **$299** | $399 | $499 | $599 | |

Besides brand and price, we see that the attributes were defined in terms of features, including the vacuum's performance (i.e., what percentage of dirt and debris it picks up), capacity (i.e., how often it needs to be emptied), the type of navigation (i.e., does it change directions by just bumping into things or is it "smart" and able to scan and determine an optimal path), where it can be programmed, and whether or not virtual borders can be set to keep the robotic vacuum away from certain areas of the home. A summary of the data using model notation is provided in Table 4.

**Table 4. Data Summary**

| Choice Variables | Description |
|---|---|
| $N = 332$ | total number of respondents |
| $H = 16$ | number of choice tasks for each respondent $n$ |
| $P = 5$ | number of alternatives in each choice task |
| $L = 12$ | number of attribute levels in each choice task |
| **GoM Variables** | **Description** |
| $J = 18$ | number of categorical questions |
| $n_j = 2$ | number of categorical responses for each question $j$ |

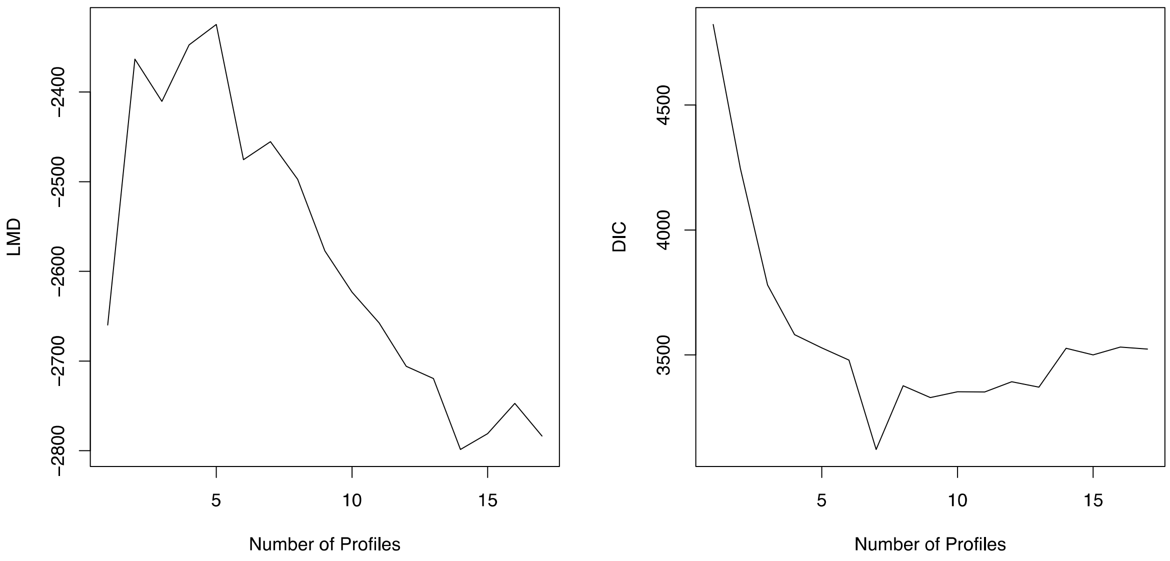
## 4 Results

We report the results of three models. The Intercept model only includes an intercept in the upper level model (i.e., $\beta_n = \gamma + \xi_n$) and serves as a baseline. The Binary Covariates model includes all 18 dummy-coded statements from Table 2 as covariates in the upper level model (i.e., $\beta_n = \Gamma' z_n + \xi_n$) and represents the typical way these discrete covariates would be used in practice. Finally, the Membership Vector model is our proposed model, which uses the membership vectors from the grade of membership model as covariates for $K = 5$ profiles (i.e., $\beta_n = \Gamma' g_n + \xi_n$).

The number of profiles $K$ is determined by the analyst. Following the review on model selection criteria by Joutard *et al.* (2007), we ran an isolated GoM model on the 18 statements in Table 2 and compared two measures of fit. The first is the Newton-Raftery approximation of the log marginal density (LMD) (Newton and Raftery 1994), a standard Bayesian measure. The second is the deviance information criterion (DIC), developed by Spiegelhalter *et al.* (2002). Values closer to zero indicate improvement in fit for both measures. Figure 7 includes charts for the values of both LMD and DIC for models with $K$=2 to $K$=18. According to the LMD (where values closer to zero indicate better fit), $K$=5 is best. According to the DIC (where values closer to zero indicate better fit), $K$=7 is best. With the range of possible models narrowed, we ran the proposed model for $K = 5$ to $K = 7$. Comparing results to find profiles that are sufficiently differentiated and non-repeating, the model with $K = 5$ was deemed best.

The final 75 respondents were reserved as a holdout sample, leaving 257 respondents for calibration. In addition, one choice task was held out from each respondent in the calibration sample for an additional measure of predictive fit. We ran each model for 50,000 iterations, saving every 50th draw, and using the final 20,000 iterations for inference. We checked for but found no substantial evidence of label switching.

**Figure 7. Selecting the Number of K**



Choice model LMD is used for in-sample fit. Out-of-sample fit is provided in terms of hit probabilities. A hit probability is the average posterior probability of a set of observed choices given a specific model. The hit probability is averaged over a set of respondents, observations, and post-burn-in MCMC draws. The two hit probabilities of interest are for the holdout tasks from the calibration sample and the holdout sample, respectively. For the calibration holdout task hit probability, $N$ = 257, $H$ = 1, $R$ = 401, and the part-worth draws are available from each model. For the holdout sample hit probability, $N$ = 75, $H$ = 16, $R$ = 401, and the part-worths are drawn from the distribution of heterogeneity. However, the covariates in the proposed model are generated as part of the model and thus are not available for the holdout sample.

To address this, the observed choices for the respondents in the holdout sample were withheld while their observed categorical responses were included to produce the covariates needed to compute the hit probability. Following Gelman *et al.* (2013), we treat the withheld observed choices for the holdout sample respondents as missing data and employ data augmentation to impute the missing observations at each iteration in the MCMC chain. This allows us to produce covariates for the holdout sample that are informed by the complete model, including the holdout sample's observed categorical responses and the calibration sample's observed choices and categorical responses, and thus draw the part-worths to compute the hit probability. We perform this data augmentation for the holdout sample respondents for each of the reported models.

**Table 5. Model Fit**

|  | In-Sample | Out-of-Sample |  |
|---|---|---|---|
| Model | LMD | Hit Prob.[1] | Hit Prob.[2] |
| Intercept $\beta_n = \gamma + \xi_n$ | -2441.048 | 0.654 | 0.371 |
| Binary Covariates $\beta_n = \Gamma' z_n + \xi_n$ | -2424.537 | 0.651 | 0.300 |
| Membership Vector $\beta_n = \Gamma' g_n + \xi_n$ | -2307.619 | 0.670 | 0.451 |

[1] Using calibration respondent hold-out tasks.

[2] Using hold-out sample respondents where $\beta^M_{n^*,r^*} \sim N(\Gamma^{M\prime}_{r^*} z^M_n, V^M_{\beta,r^*})$ or $N(\Gamma^{M\prime}_{r^*} g^M_{n^*,r^*}, V^M_{\beta,r^*})$.

Table 5 demonstrates that, across all measures of model fit, covariates uncovered with mixed membership modeling have more explanatory and predictive power than standard models using discrete covariates. Another alternative to the proposed model would be to include interactions directly. However, in running this alternative model, problems manifested themselves with only two-way interactions. First, the flexibility of the model induced by including so many covariates clearly allowed for overfitting. As we increased the number of iterations in the Markov chain, we continued to see an improvement in in-sample fit with no change in predictive fit and no sign of convergence. Second, the number of interactions would make interpretation infeasible. For these reasons we don't report the results of this model.

The proposed model also improves inference regarding the drivers of preference heterogeneity. To illustrate, let's consider the posterior means of Γ from the Binary Covariates model. Table 6 displays the complete Γ matrix. The attribute levels are on the left and each column in the matrix is associated with the intercept or one of the statements from Table 2. The posterior means in bold are more than two standard deviations below or above zero. This matrix should inform a marketer concerning the drivers of preference for promotion and targeting strategies. However, making sense of the significant values or considering how these items may interact is cumbersome.

**Table 6** Binary Covariates Model Γ Estimates

| Attribute Levels | Int. | S1 | S2 | S3 | S4 | S5 | S6 | S7 | S8 | S9 | S10 | S11 | S12 | S13 | S14 | S15 | S16 | S17 | S18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Neato | 1.52 | 0.03 | 0.22 | -0.05 | -0.31 | **2.19** | -1.36 | -0.78 | **-2.18** | -0.74 | -0.29 | **3.75** | -1.03 | 1.65 | -0.83 | -0.66 | 0.56 | -1.25 | -0.17 |
| iRobot | **2.76** | 0.26 | 0.79 | -0.31 | -0.50 | **2.41** | -1.73 | -1.65 | **-2.53** | -0.26 | -0.26 | 3.27 | -0.84 | 1.80 | -0.95 | -0.24 | 0.52 | -0.69 | 0.22 |
| Samsung | **2.96** | -0.46 | 0.42 | -0.19 | -0.23 | **2.27** | -1.53 | -1.02 | -1.82 | -0.76 | -0.39 | 3.69 | -0.95 | 1.37 | -1.38 | -0.83 | 1.00 | -0.77 | -0.27 |
| Black & Decker | 2.69 | -0.42 | 0.21 | -0.53 | -0.10 | **1.99** | -0.86 | -1.70 | -1.85 | -0.55 | -0.05 | **4.23** | **-1.81** | **2.77** | -0.60 | -0.58 | 0.82 | -1.18 | -0.07 |
| Performance: 85% | 0.91 | 0.43 | 0.29 | 0.47 | 0.68 | **-1.17** | 0.10 | 0.93 | 0.68 | -0.25 | 0.21 | -0.43 | **0.92** | -0.31 | 0.83 | 0.34 | 0.92 | -0.15 | 0.61 |
| Capacity: Every 2-3 uses | 0.47 | -0.22 | 0.24 | 0.09 | 0.26 | -0.02 | 0.21 | 0.07 | 0.03 | -0.36 | -0.15 | -0.07 | 0.02 | 0.19 | -0.10 | 0.14 | 0.43 | -0.19 | -0.14 |
| Smart Navigation | 0.25 | 0.47 | -0.32 | 0.30 | 0.14 | 0.21 | 0.40 | 0.18 | 0.06 | 0.21 | 0.11 | -0.34 | -0.21 | -0.14 | -0.14 | -0.22 | **-0.74** | 0.31 | -0.38 |
| App Programming | -0.50 | 0.14 | 0.48 | -0.07 | -0.12 | **0.57** | -0.07 | -0.10 | -0.12 | **0.63** | 0.22 | **-1.15** | 0.05 | 0.34 | -0.29 | 0.16 | -0.38 | -0.04 | -0.21 |
| Virtual Borders | **1.01** | -0.22 | -0.29 | 0.23 | 0.23 | -0.10 | 0.07 | -0.40 | 0.12 | 0.02 | -0.03 | -0.18 | -0.08 | **-1.18** | 0.30 | 0.21 | -0.63 | 0.04 | 0.44 |
| $399 | -1.10 | -0.22 | 0.07 | -0.36 | -0.40 | 0.43 | 0.82 | **-1.11** | 0.73 | **-1.43** | 0.03 | **1.90** | **-1.16** | -0.71 | -0.07 | 0.65 | -0.22 | -0.28 | -0.41 |
| $499 | **-3.04** | -0.87 | 0.37 | 0.13 | -1.29 | 1.22 | 1.14 | **-1.97** | 1.02 | **-2.54** | -0.30 | **3.22** | **-2.93** | -1.48 | 0.07 | 1.57 | -0.87 | 0.33 | -0.60 |
| $599 | **-4.83** | -0.96 | 0.81 | 0.61 | -1.41 | 0.77 | 0.76 | -1.75 | 0.90 | -2.69 | -0.29 | **4.28** | **-3.41** | -2.56 | -0.71 | 2.04 | -1.64 | 0.16 | -0.43 |

For example, we can use Table 6 to infer that respondents who are concerned about germs and dirt (i.e., statement 5 “I worry about germs and dirt on my floor and carpet”) prefer any brand of robotic vacuum relative to the outside good while not being concerned about getting the highest level of performance. We might expect this is because they are cleaning frequently (e.g., statement 10 “I spend over two hours per week cleaning”) and having a robotic vacuum is simply one part of a larger cleaning solution. Without a way to properly account for interactions, we aren’t able to understand these more detailed explanations of preference heterogeneity.

The proposed model accounts for such interactions by identifying differentiated respondent profiles. Table 7 details the profiles as described by the estimates of $\lambda_{j,k}(1)$. Since the respondents were qualified by owning or being interested in a robotic vacuum, it isn’t surprising that every profile has statement 1 “I enjoy coming home to a clean house” occurring with high probability. Profile 1 is differentiated from the other models by statement 2 “I don’t feel relaxed when I know my home isn’t clean,” statement 10 “I spend over two hours per week cleaning,” and statement 5 “I worry about germs and dirt on my floor and carpet” occurring with high probability and statement 11 “I have a cleaning person who cleans for me” occurring with the lowest probability. We name this profile “Constantly Cleaning.”

Profile 2 is differentiated by statement 12 “Robotic vacuums are too expensive,” statement 9 “I don’t spend much time cleaning,” and statement 10 “I spend over two hours per week cleaning” occurring with high probability and statement 13 “Robotic vacuums are too complicated to program, set up, and operate” occurring with the lowest probability. We name this profile “Price Sensitive with Little Cleaning.” Profile 3 is differentiated by statement 2 “I don’t feel relaxed when I know my home isn’t clean,” statement 7 “I don’t like going to someone’s home that is dirty,” and statement 6 “I get anxious about having guests when my home is dirty” occurring with high probability. We name this profile “Anxious about Cleanliness.”

Profile 4, like profile 2, has statement 12 “Robotic vacuums are too expensive” occurring with high probability, but is further differentiated by statement 4 “I have trouble keeping the floor beneath my furniture clean” and statement 14 “Robotic vacuums often need to be ‘rescued’ because they get stuck.” We name this profile “Price Sensitive with Difficulty Cleaning.” Finally, profile 5, like profile 3, has statements 6, 7, and 2 occurring with high probability—statements describing being anxious about cleanliness—as well as, like profile 4, a high probability of statements 14 and 4, which describe difficulty cleaning along with a belief that robotic vacuums get stuck. We name this profile “Anxious and Suspicious.”

**Table 7. Membership Vector Model $\lambda_{j.k}(1)$ Estimates**

| No. | Statements | $\lambda_{j,1}(1)$ | $\lambda_{j,2}(1)$ | $\lambda_{j,3}(1)$ | $\lambda_{j,4}(1)$ | $\lambda_{j,5}(1)$ |
|---|---|---|---|---|---|---|
| 1 | I enjoy coming home to a clean house. | 0.75 | 0.65 | 0.87 | 0.89 | 0.96 |
| 2 | I don't feel relaxed when I know my home isn't clean. | 0.56 | 0.14 | 0.82 | 0.48 | 0.89 |
| 3 | I worry about pet hair and dander in the home. | 0.36 | 0.14 | 0.58 | 0.47 | 0.82 |
| 4 | I have trouble keeping the floor beneath my furniture clean. | 0.28 | 0.14 | 0.44 | 0.67 | 0.83 |
| 5 | I worry about germs and dirt on my floor and carpet. | 0.50 | 0.11 | 0.77 | 0.49 | 0.83 |
| 6 | I get anxious about having guests when my home is dirty. | 0.46 | 0.28 | 0.79 | 0.53 | 0.93 |
| 7 | I don't like going to someone's home that is dirty. | 0.19 | 0.18 | 0.80 | 0.51 | 0.91 |
| 8 | I don't like touching dirty things. | 0.16 | 0.12 | 0.75 | 0.18 | 0.87 |
| 9 | I don't spend much time cleaning. | 0.09 | 0.31 | 0.11 | 0.44 | 0.06 |
| 10 | I spend over two hours per week cleaning. | 0.51 | 0.26 | 0.65 | 0.41 | 0.87 |
| 11 | I have a cleaning person who cleans for me. | 0.07 | 0.04 | 0.14 | 0.04 | 0.08 |
| 12 | Robotic vacuums are too expensive. | 0.36 | 0.63 | 0.28 | 0.92 | 0.60 |
| 13 | Robotic vacuums are too complicated to program, set up, and operate. | 0.09 | 0.04 | 0.08 | 0.23 | 0.19 |
| 14 | Robotic vacuums often need to be "rescued" because they get stuck. | 0.21 | 0.26 | 0.25 | 0.35 | 0.86 |
| 15 | Robotic vacuums need to have their trash containers changed too often. | 0.27 | 0.17 | 0.21 | 0.15 | 0.50 |
| 16 | Robotic vacuums don't do a good enough job cleaning the floor and carpet. | 0.16 | 0.06 | 0.22 | 0.22 | 0.40 |
| 17 | Robotic vacuums don't spend enough time on the really dirty spots on the floor. | 0.15 | 0.26 | 0.18 | 0.20 | 0.18 |
| 18 | Robotic vacuums scare household pets. | 0.17 | 0.17 | 0.23 | 0.26 | 0.42 |

**Table 8. Profile Names**

| No. | Profile Names |
|---|---|
| 1 | Constantly Cleaning |
| 2 | Price Sensitive with Little Cleaning |
| 3 | Anxious about Cleanliness |
| 4 | Price Sensitive with Difficulty Cleaning |
| 5 | Anxious and Suspicious |

Table 9 displays the membership vectors to variability in the part-worths. Again, the posterior means in bold are more than two standard deviations below and above zero. Note that the size of the coefficients is in part a function of the size of $K$ and the sum-to-one constraint on $g_n$. As $K$ increases in size, each element of the membership vector $g_n$ gets smaller and the coefficients of $\Gamma$ get larger to map to the part-worth estimates. Even taking this constraint into account, the coefficients are still larger than those produced by the standard model as represented in Table 6. This is because partial membership in these extreme profiles allows the distribution of preferences to move into the extremes. Regardless, the focus in interpreting the coefficients in Table 9 remains on their relative sign and magnitude.

**Table 9. Membership Vector Model Γ Estimates**

| Attribute Levels | P1 | P2 | P3 | P4 | P5 |
|---|---|---|---|---|---|
| Neato | **26.49** | **-16.13** | 0.27 | **9.04** | **-20.17** |
| iRobot | **27.52** | **-11.49** | 1.84 | **9.86** | **-20.75** |
| Samsung | **24.28** | **-14.70** | **8.57** | **9.18** | **-21.94** |
| Black & Decker | **25.96** | **-15.93** | 4.19 | **10.59** | **-21.59** |
| Performance: 85% | -1.19 | -0.68 | -2.15 | 2.00 | **20.47** |
| Capacity: Every 2-3 uses | **1.93** | -0.62 | 0.18 | 0.19 | **1.70** |
| Smart Navigation | 0.27 | **3.10** | -1.12 | -0.01 | **2.04** |
| App Programming | **1.49** | 0.77 | -1.23 | -1.33 | -0.06 |
| Virtual Borders | 0.00 | **4.87** | -1.38 | -1.17 | 1.31 |
| $399 | 1.84 | -0.30 | **3.51** | **-16.36** | 1.07 |
| $499 | 3.26 | -0.32 | 4.70 | **-36.66** | 0.37 |
| $599 | 3.09 | -0.93 | **8.40** | **-53.50** | -2.58 |

As with Table 6, the matrix in Table 9 should inform a marketer concerning the drivers of preference for promotion and targeting strategies. However, using the proposed model, we are able to explain preferences in terms of the extreme profiles. For example, profile 1, "Anxious about Cleanliness" includes statements 5 "I worry about germs and dirt on my floor and carpet" and 10 "I spend over two hours per week cleaning" with high probability. With this profile we can answer what was only suggested from Table 6, that the more an individual is aligned with this profile, the more they prefer any brand of robotic vacuum while caring about a high-capacity robotic vacuum rather than one that performs the best. In other words, since they are cleaning often, they want a robotic vacuum with high capacity in order to effectively assist but not replace other cleaning efforts.

We can better inform targeting and promotion strategies using the proposed model. We can use the estimate of Γ as a roadmap for targeting by matching what respondents prefer with a more detailed explanation of what is driving those preferences. For example, for consumers above a certain threshold in their partial membership in profile 4 “Price Sensitive with Difficulty Cleaning,” we know that pricing promotions should be especially effective since they have a need for robotic vacuums but are incredibly price sensitive. The dimension-reduction provided by employing a GoM model makes this plausible with the 12 × 5 Γ matrix in Table 9 compared with a similar task using the 12 × 19 Γ matrix in Table 6 from the alternative model or an even larger Γ matrix that includes interactions directly.

Accounting for the co-occurrence or interactions among items is akin to segmenting the market. The blocks of significant attribute level coefficients in Table 9 are reminiscent of such segmentation solutions. Unlike mixture models, which are typical in clustering applications, where a respondent is assigned to a single category, mixed membership models like the GoM allow for the more realistic description of each respondent being a partial member of each profile. In our empirical application, it makes sense that consumers interested in robotic vacuums are not going to be constantly cleaning, anxious about cleanliness, skeptical of robotic vacuums, or price sensitive exclusively. Rather, each individual is a mix of all the profiles, with weights determined heterogeneously. Accounting for such differences improves our ability to conduct inference.

## 5 Discussion

In this paper we show that modeling interactions among discrete multivariate data does more to explain consumer preferences than the discrete covariates on their own. This is accomplished by combining a grade of membership model, part of the class of mixed membership models, with choice modeling to estimate membership vectors for use in a hierarchical Bayesian random effects distribution of heterogeneity. Note that our discrete multivariate data in this application consist of pick any/*J* data. However, it is applicable to any discrete data, including rating scale data.

Choice modeling remains an essential fixture of marketing research. However, finding covariates that are explanatory of preference heterogeneity has proven difficult. Our proposed model provides a novel way to account for interactions, and provide dimension reduction, for survey data that explain variation in part-worth utilities. The empirical application utilizes typical survey response data to demonstrate the use of the proposed model. However, with growing access to unstructured collections of discrete data, we see this approach as an important step to utilizing such data, including text, to improve choice modeling.

Latent Dirichlet allocation, as another kind of mixed membership model, performs in a similar way to the GoM. Text data results in the same kind of sparse matrix as the multinomial data used in the GoM model, with LDA proceeding with words instead of items or statements and a single document for each individual. The dimension reduction using text is even more dramatic when starting with potentially thousands of unique words in the count matrix. However, the amount of data needed to run LDA with words composing the collection of discrete data is significant due to the large number of words in any given vocabulary. Without enough data, there are a variety of developments in topic modeling that are ripe for application within marketing, including using Dirichlet process priors (Ferguson 1973, Antoniak 1974) as a kind of distribution of heterogeneity over topic proportions. We leave the practical problems of using text in the place of traditional survey questions as an extension to this research.

Another extension relates to estimating the optimal size of K. While there isn’t a consensus as to which measure of model fit provides the gold standard for determining the size of K, there are a number of extant methods for navigating across possible model dimensions that could be employed to include K as a parameter in the model (Green 1995, Green *et al*. 2015). The technical details of how to incorporate such methods into the proposed model is left for future research.

More generally, we see the use of mixed membership models as a model-based approach to classifying consumers that yields a more realistic description of the individual as being a mixture of various extreme consumer profiles. This paper serves as a step toward fulfilling a broader need to provide more complete descriptions and explanations of consumer preference heterogeneity.

Marc R. Dotson

Joachim Büschken

Greg M. Allenby

## References

Airoldi, Edoardo M, David Blei, Elena A Erosheva, Stephen E Fienberg. 2014. Handbook of Mixed Membership Models and Their Applications. 1st ed. Chapman & Hall/CRC.

Allenby, G M, James L Ginter. 1995. Using Extremes to Design Products and Segment Markets. Journal of Marketing Research 32(4) 392–403.

Allenby, Greg M, Neeraj Arora, James L Ginter. 1998. On the Heterogeneity of Demand. Journal of Marketing Research 35(3) 384–389.

Allenby, Greg M, Peter E Rossi. 1998. Marketing Models of Consumer Heterogeneity. Journal of Econometrics 89(1–2) 57–78.

Antoniak, Charles E. 1974. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. The Annals of Statistics 2(6) 1152–1174.

Archak, Nikolay, Anindya Ghose, Panagiotis G Ipeirotis. 2011. Deriving the Pricing Power of Product Features by Mining Consumer Reviews. Management Science 57(8) 1485–1509.

Blei, David M, Jon D McAuliffe. 2007. Supervised topic models. Neural Information Processing Systems.

Blei, David M, Andrew Y Ng, Michael I Jordan. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3 993–1022.

Büschken, Joachim, Greg M Allenby. 2016. Sentence-Based Text Analysis for Customer Reviews. Marketing Science (forthcoming).

Chandukala, Sandeep R, Yancy D Edwards, Greg M Allenby. 2011. Identifying Unmet Demand. Marketing Science 30(1) 61–73.

Clive, Jonathan, Max A Woodbury, Ilene C Siegler. 1983. Fuzzy and Crisp Set-Theoretic-Based Classification of Health and Disease. Journal of Medical Systems 7(4) 317–332.

Erosheva, Elena A. 2002. Grade of Membership and Latent Structure Models with Application to Disability Survey Data. Ph.D. thesis, Department of Statistics, Carnegie Mellon University.

Erosheva, Elena A, Stephen E Fienberg, Cyrille Joutard. 2007. Describing Disability Through Individual-Level Mixture Models for Multivariate Binary Data. The Annals of Applied Statistics 1(2) 346–384.

Ferguson, Thomas S. 1973. A Bayesian Analysis of Some Nonparametric Problems. The Annals of Statistics 1(2) 209–230.

Galyardt, April. 2014. Interpreting Mixed Membership. Handbook of Mixed Membership Models and Their Applications. Chapman & Hall/CRC, 39–65.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, Donald B Rubin. 2013. Bayesian Data Analysis. Third edition ed. Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.

Gelman, Andrew, Iain Pardoe. 2006. Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models. Technometrics 48(2) 241–251.

Green, Peter J. 1995. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika 82(4) 711–732.

Green, Peter J, Krzysztof Latuszynski, Marcelo Pereyra, Christian P Robert. 2015. Bayesian Computation: A Summary of the Current State, and Samples Backwards and Forwards. Statistics and Computing 25(4) 835–862.

Gross, Justin H, Daniel Manrique-Vallier. 2014. A Mixed-Membership Approach to the Assessment of Political Ideology from Survey Responses. Handbook of Mixed Membership Models and Their Applications. Chapman & Hall/CRC, 119–139.

Horsky, Dan, Sanjog Misra, Paul Nelson. 2006. Observed and Unobserved Preference Heterogeneity in Brand-Choice Models. Marketing Science 25(4) 322–335.

Johnson, Valen E, James H Albert. 2006. Ordinal Data Modeling. Springer Science & Business Media.

Joutard, Cyrille, Edoardo M Airoldi, Stephen E Fienberg, Tanzy M Love. 2007. Discovery of Latent Patterns with Hierarchical Bayesian Mixed-Membership Models and the Issue of Model Choice. Data Mining Patterns: New Methods and Applications. IGI Global, Hershey, PA, USA, 1–36.

Kamakura, Wagner A, Gary J Russell. 1989. A Probabilistic Choice Model for Market Segmentation and Elasticity Structure. Journal of Marketing Research 26(4) 379–390.

Lee, Sik-Yum. 2007. Structural Equation Modeling: A Bayesian Approach, vol. 711. John Wiley & Sons.

Lee, Thomas Y, Eric T Bradlow. 2011. Automated Marketing Research Using Online Customer Reviews. Journal of Marketing Research 48(5) 881–894.

Lenk, Peter J, Wayne S DeSarbo, Paul E Green, Martin R Young. 1996. Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. Marketing Science 15(2) 173–191.

Manton, Kenneth G, Max A Woodbury, H Dennis Tolley. 1994. Statistical Application Using Fuzzy Sets. Wiley, New York.

Marini, Margaret Mooney, Xiaoli Li, Pi-Ling Fan. 1996. Characterizing Latent Structure: Factor Analytic and Grade of Membership Models. Sociological Methodology 26 133–164.

Netzer, Oded, Ronen Feldman, Jacob Goldenberg, Moshe Fresko. 2012. Mine Your Own Business: Market-Structure Surveillance Through Text Mining. Marketing Science 31(3) 521–543.

Newton, Michael A, Adrian E Raftery. 1994. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. Journal of the Royal Statistical Society. Series B (Methodological) 56(1) 3–48.

Rossi, Peter E, Greg M Allenby. 2003. Bayesian Statistics and Marketing. Marketing Science 22(3) 304–328.

Rossi, Peter E, Greg M Allenby, Robert E McCulloch. 2005. Bayesian Statistics and Marketing. J. Wiley and Sons.

Rossi, Peter E, Robert E McCulloch, Greg M Allenby. 1996. The Value of Purchase History Data in Target Marketing. Marketing Science 15(4) 321–340.

Spiegelhalter, David J, Nicola G Best, Bradley P Carlin, Angelika Van Der Linde. 2002. Bayesian Measures of Model Complexity and Fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64(4) 583–639.

Stewart, David W. 1981. The Application and Misapplication of Factor Analysis in Marketing Research. Journal of Marketing Research 18(1) 51–62.

Tanner, Martin A, Wing Hung Wong. 1987. The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association 82(398) 528–540.

Tirunillai, Seshadri, Gerard J Tellis. 2014. Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. Journal of Marketing Research 51(4) 463–479.

Woodbury, Max A, Jonathan Clive, Arthur Garson Jr. 1978. Mathematical Typology: A Grade of Membership Technique for Obtaining Disease Definition. Computers and Biomedical Research 11(3) 277–298.