

PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

August 2018

Copyright 2018

All rights reserved. No part of this volume may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from
Sawtooth Software, Inc.

FOREWORD

These proceedings are a written report of the twentieth Sawtooth Software Conference, held in Orlando, Florida, March 7-9, 2018. One-hundred seventy attendees participated. This conference has quite a long history, with the first Sawtooth Software Conference held over 30 years ago in 1987.

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 choice/conjoint analysis, MaxDiff, optimization searches for optimal product lines, key drivers analysis, and market segmentation and classification.

The papers and discussant comments are in the words of the authors and very little copyediting was performed. At the end of each of the papers are photographs of the authors and co-authors.

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

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

We are also especially grateful to the efforts of our steering committee who for many years now have helped this conference be such a success: Christopher Chapman, Keith Chrzan, Elea Feit, Joel Huber, and David Lyon.

Sawtooth Software

August, 2018

CONTENTS

CONSTRUCTED, AUGMENTED MAXDIFF.....	1
<i>Eric Bahna and Chris Chapman, Google Cloud</i>	
SHAPLEY VALUES: EASY, USEFUL AND INTUITIVE	13
<i>David W. Lyon, Aurora Market Modeling, LLC</i>	
FDA SEEKS PATIENT PREFERENCE INFORMATION TO ENHANCE THEIR BENEFIT-RISK ASSESSMENTS: CASE STUDIES	37
<i>Leslie Wilson, University of California, San Francisco and Jordan Louviere, University of South Australia</i>	
A DIRECT COMPARISON OF DISCRETE CHOICE AND ALLOCATION CONJOINT METHODOLOGIES IN THE HEALTHCARE DOMAIN	49
<i>James Pitcher, Tatiana Koudinova, and Daniel Rosen, GfK</i>	
A META-ANALYSIS ON THREE DISTINCT METHODS USED IN MEASURING VARIABILITY OF UTILITIES AND PREFERENCE SHARES WITHIN THE HIERARCHICAL BAYESIAN MODEL.....	69
<i>Jacob Nelson, Edward Paul Johnson, and Brent Fuller, Research Now—Survey Sampling International</i>	
PREFERENCE BASED CONJOINT: CAN IT BE USED TO MODEL MARKETS WITH MANY DOZENS OF PRODUCTS	87
<i>Jeroen Hardon and Marco Hoogerbrugge, SKIM Group</i>	
DEVELOPMENT OF AN ADAPTIVE TYPING TOOL FROM MAXDIFF RESPONSE DATA.....	103
<i>Jay Magidson, Statistical Innovations and John P. Madura, University of Connecticut and Statistical Innovations</i>	
COMMENTS ON “DEVELOPMENT OF AN ADAPTIVE TYPING TOOL FROM MAXDIFF RESPONSE DATA”	135
<i>Thomas C. Eagle, Eagle Analytics of California</i>	
EXTENDING THE ENSEMBLE	139
<i>Curtis Frazier, Ana Yanes, and Michael Patterson, Radius Global Market Research</i>	
SYNERGISTIC BANDIT CHOICE (SBC) DESIGN FOR CHOICE-BASED CONJOINT	149
<i>Bryan Orme, Sawtooth Software</i>	
OPTIMAL PRODUCT DESIGN BY SEQUENTIAL EXPERIMENTS	167
<i>Mingyu Joo, UC Riverside, Michael L. Thompson, Procter and Gamble, and Greg Allenby, Ohio State University</i>	
SEGMENTATION ANALYSIS VIA NON-NEGATIVE MATRIX FACTORIZATION.....	179

Michael Patterson, Jackie Guthart, and Curtis Frazier, Radius Global Market Research

VARIABLE SELECTION FOR MBC CROSS-PRICE EFFECTS 189

Katrin Dippold-Tausendpfund and Christian Neuerburg, GfK

**ACCOMMODATING MULTIPLE DATA PATHOLOGIES IN CONJOINT STUDIES VIA
CLEVER RANDOMIZATION AND ENSEMBLING 201**

*Jeffrey P. Dotson, Brigham Young University, Roger A. Bailey, The Ohio State University,
and Marc R. Dotson, Brigham Young University*

TOOLS FOR DEALING WITH CORRELATED ALTERNATIVES..... 211

Kevin Lattery and Jeroen Hardon, SKIM Group

PREDICTIVE ANALYTICS WITH REVEALED PREFERENCE/STATED PREFERENCE MODELS 225

Peter Kurz, KANTAR TNS and Stefan Binner, bms marketing research + strategy

**THE PERILS OF IGNORING UNCERTAINTY IN MARKET SIMULATIONS
AND PRODUCT LINE OPTIMIZATION..... 237**

Scott Ferguson, North Carolina State University

PROPERTIES OF DIRECT UTILITY MODELS FOR VOLUMETRIC CONJOINT 257

Jake Lee, Quantum Strategy, Inc

A COMPARISON OF VOLUMETRIC MODELS 267

*Thomas C. Eagle, Eagle Analytics of California, Inc., Jordan Louviere,
University of South Australia, and Towhidul Islam, University of Guelph, Canada*

DIRECT ESTIMATION OF KEY DRIVERS FROM A FITTED BAYESIAN NETWORK 293

Benjamin Cortese, KS&R

PRODUCT RELEVANCE AND NON-COMPENSATORY CHOICE..... 313

*Marc R. Dotson, Brigham Young University, Roger A. Bailey, and Greg M. Allenby,
The Ohio State University*

SUMMARY OF FINDINGS

The twentieth Sawtooth Software Conference was held in Orlando, Florida, March 7-9, 2018. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2018 Sawtooth Software Conference Proceedings.

* **Constructed, Augmented MaxDiff** (Eric Bahna and Chris Chapman, Google Cloud): The authors described how they have used multiple MaxDiff surveys to assess customer prioritization of potential features and usage scenarios at Google Cloud. However, the surveys have been long and involved, asking respondents to evaluate many items that were not relevant to them or that were unimportant to them. The authors created an adaptive MaxDiff survey that first asked respondents to mark any items that were irrelevant or unimportant. The remaining items were moved forward to a subsequent MaxDiff exercise (plus a few irrelevant and unimportant items were also carried forward, to assist with issues of scaling). This was made possible in survey programming by using the constructed list function in Lighthouse Studio. Analysis was conducted in R. The data were augmented by adding tasks indicating that the unimportant (and dropped) items were less preferred than the relevant and included items for each respondent. Eric and Chris found that the data appeared to be of improved quality (because respondents were giving more input on the items that they saw as relevant to them) and respondents reported a better experience. They indicated that there are some open questions regarding how much augmentation affects the utility results depending on the pattern of screening answers, as well as questions regarding the details involved in the composition of the constructed lists.

* Best Presentation based on audience voting.

Shapley Values: Easy, Useful and Intuitive (David W. Lyon, Aurora Market Modeling, LLC): Shapley Values are a widely useful way to summarize how much individual items contribute to the overall value of combinations of items (e.g., product assortments, feature bundles, sets of advertising claims). Although it often is applied within TURF (Total Unduplicated Reach & Frequency) simulations, David explained that Shapley Value is not limited to TURF. It can be applied to regression analysis and many other types of analyses. David argued that Shapley Values are easier to think about, easier to present and more memorable than the endless lists of combinations reported by TURF. While clients often ask for the one best combination, he believes that their true underlying objective is likely to be an understanding or insight into item values or patterns. One appealing aspect of Shapley Values is that they are far more stable than analyses relying on individual combinations or orders. They are far more stable in the face of minor data changes, meaning that changes in results are far more likely to be real. David provided R code and described the challenges and potential solutions for dealing with the computational complexity of computing Shapley Values.

FDA Seeks Patient Preference Information to Enhance Their Benefit-Risk Assessments: Case Studies (Leslie Wilson, University of California, San Francisco and Jordan Louviere, University of South Australia): In 2013, the FDA launched the Patient Preference Initiative to incorporate patients' views as scientific, empirical evidence when appropriate. Due to the strengths of conjoint analysis (e.g. CBC) to measure respondent preferences, the FDA now accepts conjoint analysis (and best-worst analysis) to help in the regulatory approval of medical devices. The authors presented a few examples of case studies to help the audience gain insights into the growth of patient preference research within the FDA and to gain some appreciation of

the nuances of using CBC in this field. Medical devices often involve potential life-threatening risks, and the use of CBC allows researchers to have respondents trade off and thereby quantify the amount of risk they are willing to accept to gain often life-saving benefits. This allows the FDA to make appropriate decisions involving risk tradeoffs when approving medical devices that can benefit society as a whole.

A Direct Comparison of Discrete Choice and Allocation Conjoint Methodologies in the Healthcare Domain (James Pitcher, Tatiana Koudinova, and Daniel Rosen, GfK): GfK commonly uses two distinct methodologies to estimate new product preference shares in the healthcare and pharma space: 1) patient-based discrete choice asks physicians to report prescribing preferences for specific real world patients, and 2) allocation-based conjoint asks physicians to report their prescribing decisions at a practice level (allocation of patients) rather than on a per-patient basis. This distinction reflects a trade-off market researchers often make when designing a research study: whether to have the research environment closely resemble the real world decision environment, or whether a carefully designed, albeit “artificial,” research environment elicits more accurate information from respondents. The authors also tested an approach for encouraging physician respondents to provide more realistic answers, by assigning the respondents a letter grade after the conjoint exercise depending on the respondents’ internal consistency. The authors believed that physicians (who completed many years of schooling) would be positively motivated by the consistency grading system. The authors noted significant differences in the predictions from the allocation-based vs. per-patient based questionnaires. With the allocation-based format, some respondents treated the allocation as if the values needed to sum to 100%, even though it was stressed to them that multi-therapy prescriptions would mean values summing to larger than 100%. With the patient-based approach, the authors felt that respondents likely were able to recall more extreme patients which could bias the survey responses.

A Meta-Analysis on Three Distinct Methods Used in Measuring Variability of Utilities and Preference Shares within the Hierarchical Bayesian Model (Jacob Nelson, Edward Paul Johnson, and Brent Fuller, Research Now–Survey Sampling International): Researchers using HB estimates (especially those who use Sawtooth Software tools) often collapse the individual-level draws and run simulations on the point estimates. They often use the standard frequentist approach to estimate the confidence levels by applying the common formulas for estimating standard errors. Those following the Bayesian tradition have estimated confidence intervals by using lower-level draws, upper-level draws, or simulated draws from the upper-level parameters (leveraging the means and covariances for the population). These methods produce different results in terms of characterizing uncertainty. The authors examined three approaches using 50 conjoint or MaxDiff research projects conducted by Survey Sampling. For CBC studies, the authors found that the point-estimate method understates the uncertainty in most every case compared to the other two methods. The authors computed a ratio between the widths of 95% credible and confidence intervals for raw part-worths between methods for CBC data, finding that the lower-level posterior distribution’s credible interval is, on average, 2.2 times larger than the point estimate’s confidence interval, and the upper level posterior distribution method’s credible interval is, on average, 2.9 times larger. They also demonstrated that the mean share predictions from CBC data can be different between the methods. However, when confidence intervals for shares of preference were estimated, the point-estimate method gave closer results to the two methods of simulating on the draws. With the MaxDiff datasets and raw utility parameters, the authors found much greater similarity in the estimated confidence intervals

among the three tested methods. However, when the raw parameters were converted to MaxDiff shares of preference, the point estimate method of computing uncertainty reported larger uncertainty bands than the two Bayesian methods (opposite of what was found for CBC studies).

Preference-Based Conjoint—Can It Be Used to Model Markets with Many Dozens of Products? (Jeroen Hardon, Marco Hooggerbrugge, SKIM Group): The authors presented two conjoint approaches that aim to predict better in situations when one has dozens of products in the simulator, but only a few concepts shown per task during the interview. They called the two approaches PBC (Preference-Based Conjoint) and PBC-squared. Both techniques oversample levels that respondents are believed to prefer, so they are adaptive techniques. Jeroen and Marco designed multiple experiments to test their new approaches versus the existing CBC and ACBC approaches. Their findings suggest that both PBC approaches might be an improvement over current practices, but indicated that more work is still needed to refine and confirm their findings.

Development of an Adaptive Typing Tool from MaxDiff Response Data (Jay Magidson, Statistical Innovations, and John P. Madura, University of Connecticut and Statistical Innovations): The authors demonstrated a new adaptive approach for developing MaxDiff typing tools that achieves high accuracy with only 8 binary comparisons (tasks, pairs) in an 8-segment example. Reduction to 7 tasks can be achieved if triples are included in the mix. Jay and John provided a theoretical framework for further task reduction by applying a hierarchical latent class tree (LCT) structure to reduce segment similarity. Preliminary results with and without adjustment for scale confounds suggest that the LCT approach not only yields further task reduction but also provides more meaningful segments. The authors also emphasized that a direct 1-step approach to developing segments via latent class MNL is preferred over a 2-step HB followed by clustering.

Extending the Ensemble (Curtis Frazier, Ana Yanes, and Michael Patterson, Radius Global Market Research): The use of ensembles (multiple solutions) in cluster analysis has been a positive development for marketing scientists. The authors investigated the value of not only varying the clustering algorithms used in building the ensemble, but in varying the basis variables included as inputs to the clustering routines. Using synthetic data, they compared the relative performance of this approach compared to existing approaches. They found that varying the inputs in general did not appear to be significantly better than ensemble methods that vary the clustering algorithms and the number of groups in the candidate solutions (except in cases with extreme skews in segment sizes). The authors wondered, however, if there was reason to believe that the structure of their synthetic data may have been inhibiting the ability for this approach (randomly omitting a subset of basis variables) to shine.

Synergistic Bandit Choice (SBC) for Choice-Based Conjoint (Bryan Orme, Sawtooth Software): When studying concepts involving strong higher-order interaction effects, Synergistic Bandit Choice (SBC) for Choice-Based Conjoint can perform better than standard CBC studies. Bryan suggested that SBC is most useful for situations such as developing FMCG concepts involving aesthetic packaging (style, color, claims, packaging graphics, brand name, nutritional content, etc.) where it is expected that there may be strong and complex higher-order interaction effects (among 3+ attributes at a time) that are difficult to measure using traditional CBC design strategies. SBC leverages the collective knowledge of previously interviewed respondents, filters their choices to focus on the most significant interaction effects, and then oversamples the most synergistic feature combinations for evaluation by subsequent respondents. Bryan showed

results for a pilot study that demonstrated the value of the new technique for a conjoint design involving very strong high-order interaction effects.

Optimal Product Design by Sequential Experiments (Mingyu Joo, UC Riverside, Michael L. Thompson, Procter & Gamble, Greg M. Allenby, Ohio State University): Optimal product and package design relies on identifying interactions among attributes and their levels. Product and package colors, tag lines, styles and visuals are examples of attributes with a “flat” space that is difficult to parameterize. Compounding this problem is the interest to identify interactive effects among attribute-levels, such as certain color combinations and messaging strategies that are thought to increase sales. The authors presented a general framework for identifying these high-dimensional interactions in the context of a sequential experiment. Their proposed design criterion differs from traditional experiment design criteria, such as D-optimality, which seeks to minimize the variance of all model parameters. Their design criterion favors product configurations with a high likelihood of improving upon the best configuration already tested. They demonstrated their model within the context of a package design problem faced by a leading consumer packaged goods manufacturer. Through simulation, they found that five rounds of a sequential experiment were sufficient for interaction detection. They designed a study and applied their model to data in which respondents identify the best package design from a set of alternatives. The best package design from their sequential experiment was then compared to alternative designs that were deemed best from alternative methodologies in a second study. The results from this best-of-class comparison favored their proposed method in comparison to other methods used.

Segmentation Analysis via Non-Negative Matrix Factorization (Michael Patterson, Jackie Guthart, and Curtis Frazier, Radius Global Market Research): Non-Negative Matrix Factorization (NMF) is a relatively new technique that allows for the simultaneous segmentation of individuals and “factoring” of variables. The authors demonstrated that NMF performs very well, especially in the case of highly correlated datasets. An NMF analysis simultaneously takes into account the relationship between the segmentation basis variables while also forming the segments. That is, items are grouped together in “factors” or latent variables, at the same time that individual respondents are grouped together in segments. In the case of some highly correlated datasets that the authors submitted to different segmentation procedures, NMF significantly outperformed the other methods. Like any other method, NMF comes with its limitations. For one, one is limited to using it only on non-negative datasets.

Variable Selection for MBC Cross-Price Effects (Katrin Dippold-Tausendpfund and Christian Neuerburg, GfK): In Menu-Based Choice experiments, cross-price effects need to be selected carefully to avoid overfitting the models or having simulation results distorted by “noisy” parameters. The authors investigated different approaches that support the selection of cross-price effects and compared their performance based on synthetic datasets under varying data conditions. They found that selection approaches that result in sparse models, e.g., variable selection with lasso, do very well under different data settings, especially with respect to the KPI measure that quantifies the quality of the resulting pricing decision. But also the Relationship Chi-squared test that is statistically less advanced and already implemented in the MBC software performs very well if the p-value cut is selected carefully. The authors emphasized that complex choice menus require a strict variable selection.

Accommodating Multiple Data Pathologies in Conjoint Studies via Clever Randomization and Ensembling (Jeffrey P. Dotson, Brigham Young University, Roger A. Bailey, The Ohio State University, and Marc R. Dotson, Brigham Young University): The authors described how ensemble-based approaches currently dominate the world of competitive out-of-sample prediction. From Kaggle to the Netflix Prize, the predictive power inherent in using many models overshadows prediction reliant on the performance of a single model. The primary reason ensembles predict so well is that they serve as a hedge against model misspecification. Since we have uncertainty about the correct model for any given context, running many models and producing a consensus is a simple yet powerful way to improve predictions. The authors demonstrated a simple approach to generating ensembles from a single HB-MNL model that improved holdout predictions both for simulated and real conjoint data sets. The approach involved looping over each respondent's HB-MNL utilities and for each iteration randomly selecting a subset of attributes for the respondent to ignore (by setting their utilities to zero) and a subset of attribute levels for the respondent to screen out (by setting their utilities to negative infinity). Consensus (modal) predictions of holdouts across iterations were calculated for each respondent and the predictions were better than predictions from a single traditional conjoint analysis model.

Tools for Dealing with Correlated Alternatives (Kevin Lattery and Jeroen Hardon, SKIM Group): Kevin and Jeroen described how IIA (Independence from Irrelevant Alternatives) can be problematic for modeling FMCG problems involving many SKUs. Individual-level estimation via HB helps reduce the IIA problems considerably in terms of aggregate simulations, though IIA still occurs within the individual-level model. Kevin and Jeroen compared approaches for resolving the IIA issues and improving the patterns of sourcing within logical nests of similar SKUs. The key comparisons they made were among standard HB, nested logit, and a post hoc nested simulation approach that is accessible to practitioners since it can be done in Excel when constructing a market simulator. They found the best results for the nested logit approach, though it is the least accessible for the typical practitioner. The other two approaches had many good properties. If using the post hoc nesting approach, they emphasized the need for creating nests that are consistent with the correlation structure of preferences evident in the data.

Predictive Analytics with Revealed Preference/Stated Preference Models (Peter Kurz, Kantar TNS, and Stefan Binner, BMS Marketing Research + Strategy): Peter and Stefan described two types of data we can employ to make market predictions in FMCG markets: Stated Preference (SP) and Revealed Preference (RP) data. SP models (like CBC surveys) have their strength in simulating dynamic markets, whereas RP models (sales, advertising, social research data) are usually preferred in static markets. The strength of combining RP-SP in models is that ability to add accurate information from the past with the power of dynamic simulations. The authors stated that the application of Time Series Corrections (RP) on Share of Choice Simulations (SP) can have a big impact, especially if one derives revenue or profit predictions. Compared to the simulation based on the DCM only (share of choice), the authors showed that corrections for trend and seasonality (RP) can lead to an improved prediction of revenue and profit. If there are no data or resources for RP-SP models, the authors recommended that one should nevertheless consider the possible impact of the point in time the study was conducted. Point in time can have a significant impact on the predictions if one ignores time series information.

The Perils of Ignoring Uncertainty in Market Simulations and Product Line

Optimization (Scott Ferguson, North Carolina State University): Scott reviewed how conjoint analysis data together with market simulators have been useful in the field of industrial engineering design. The data may be applied within optimization choice simulators to formulate effective product line strategies. In this paper, Scott explained how tuning the optimization goal to avoid worst-case outcomes (e.g. revenue or profit) can reduce the uncertainty in the real-world outcomes. Specifically, Scott demonstrated how to use the uncertainty captured in lower-level HB draws to better explore the range of possible outcomes on the choice simulator's objective function—and to avoid solutions that can lead to especially poor outcomes. An approach to avoid solutions with potentially poor outcomes is to set the maximum worst-case outcome as an objective in the multi-objective search algorithm. Scott also discussed the value of considering whether respondents are switching between products within a manufacturer's product line. Although such switching could leave overall revenue unchanged, switching behavior could lead to increased expenses for the firm (because of resource allocation decisions) or out of stock situations which can damage the firm's profitability. Therefore, switching behavior uncertainty within the firm's line could be penalized in a multi-objective search function.

Properties of Direct Utility Models for Volumetric Conjoint (Jake Lee, Quantum Strategy, Inc.): Jake commented that volumetric conjoint models are an exciting, new area for choice modeling practitioners. The new models are based on established economic theory and don't require duct tape. The models are very new and still need investigation to understand the circumstances when they work well and when adjustments need to be made. Direct utility models are more appropriate when you'd expect consumers to pick multiple options to maximize their utility. Jake described how he used a new R package to do the modeling called VDMDU, by Hardt. The model is based on Direct Utility Theory and brings in some new concepts (compared to the standard model) to help understand the consumer choice process. The two new features are the budget constraint and satiation. Volumetric models of demand for conjoint analysis are still very young, Jake asserted, though the model shows a lot of promise for managerial inference when the standard model assumptions don't fit. Jake stated that the model is a natural fit for the food and beverage categories. It could be appropriate for entertainment categories like movies and theme parks. Any time consumers would regularly pick multiple options that are competing for the consumer's budget, the direct utility model may be more appropriate than the standard model.

A Comparison of Volumetric Models (Thomas C. Eagle, Eagle Analytics of California, Jordan Louviere, University of South Australia, Towhidul Islam, University of Guelph, Canada): Volumetric models attempt to predict the number of units of a product or alternative a consumer would buy. The authors stated that volumetric models have a varied history in marketing. Some practitioners and academics avoid them, because of the complexities involved. They encouraged the audience regarding the use of volumetric models which they argued are now easier to estimate than ever before. To illustrate, they examined the patterns of substitution inherent in three different approaches to modeling volumetric data: the joint discrete/continuous model, a latent class Poisson model, and the Hardt-Allenby volumetric model. They tested the models using a volumetric choice data set involving canned tuna. The latent class Poisson model performed poorly both in terms of prediction and in terms of expected patterns of substitution/complementarity. The Hardt-Allenby model performed best in terms of prediction, but some of the substitution and complementary effects suggested by sensitivity simulations were suspect. The joint discrete/continuous model offered reasonable predictions and better face

validity in terms of the sensitivity simulations relative to what managers would typically expect about substitution and complementarity patterns. The authors concluded that the data set potentially had weaknesses and that the conclusions here are tentative, pending more evidence.

Direct Estimation of Key Drivers from a Fitted Bayesian Network (Benjamin Cortese, KS&R): Benjamin described that there are many techniques for estimating attribute-level driver scores, but the most commonly used are unable to provide information about the interactions between drivers. The introduction of Bayesian networks (BNs)—graphical representations of attribute relationships—help make sense of these complex interactions. Attempts to combine KDA and BNs through separate analysis often lead to conflicting results from the estimated top drivers and the attribute relationships depicted by the network. Benjamin proposed a new algorithm, BNKDA, to calculate driver scores directly from a fitted Bayesian network. This method relies on the Max-Min Hill-Climbing (MMHC) network fitting algorithm, Bayesian Information Criterion (BIC), and arc strengths calculated from the network. A weight factor is suggested to reduce the impact of longer paths to the target attribute. This technique provides both the directed acyclic graph (DAG) for visualizing attribute relationships and corresponding driver scores to tell a cohesive story. The algorithm was compared to two widely adopted driver analysis methods—Kruskal’s relative importance (a variant of a Shapley value) and partial least squares path modeling (PLSPM)—through simulation studies. Benjamin found that all three techniques identified similar top drivers in terms of ordering, but the magnitude of scores differed. The regression-based methods (Kruskal and PLSPM) favored directly impacting attributes in the hierarchy, while BNKDA provided more balanced estimates. He argued that consistency of driver estimates obtained from BNKDA imply that this is a viable option to calculate driver scores directly from a BN.

Product Relevance and Non-Compensatory Choice (Marc R. Dotson, Brigham Young University, Roger A. Bailey, and Greg M. Allenby, The Ohio State University): Products are composed of a variety of features or attributes. A consumer uses these attributes to infer the effectiveness of a given product to serve as a solution. The authors stated that a product that a consumer believes will be able to help address his/her specific needs and goals is relevant to that consumer; however, not all attributes are used in the same way to determine product relevance. Furthermore, the way consumers identify product relevance reveals information about the needs they want to address or the goals they seek to accomplish. Either a product’s brand is enough for a consumer to infer product relevance or the presence of certain attribute levels leads a consumer to infer product relevance. The authors developed various models that allowed them to capture these two ways to product relevance as part of an extended model of choice. Those models included conjunctive and disjunctive screening rule models. They concluded that separating and uncovering the drivers of product relevance allow firms to understand something of the underlying motivations driving consumers into the marketplace to begin with. This knowledge will help firms to design promotions and products that address those motivations, build brand loyalty, and inform consumers’ brand beliefs.

CONSTRUCTED, AUGMENTED MAXDIFF

ERIC BAHNA
CHRIS CHAPMAN
GOOGLE CLOUD

ABSTRACT

Google Cloud needed to prioritize customer needs across many product scenarios, but faced a limitation of common choice model surveys: different respondents needed to prioritize different sets of scenarios. We discuss how we solved this with Constructed, Augmented MaxDiff, and share survey design tips and R code for the method.

MOTIVATION

For Google Cloud Platform, we use MaxDiff surveys to assess customer prioritization of potential features, usage scenarios, and the like. In the course of such projects, we have encountered consistent complaints: respondents object (1) that they are unable to prioritize features that are not part of their jobs, and (2) that including all features makes a survey too lengthy and tedious. For example, one respondent commented, “[It] would be nice to have ‘no opinion’ on a particular set to not introduce noise.”

Is this just an annoyance, or is it a data quality problem? Let’s examine a hypothetical situation. Suppose we want Cloud customers to assess the importance of features related to Developer tools, No-SQL databases, and Infrastructure monitoring. Consider a respondent who is a backend developer, where infrastructure monitoring is not part of her job. If she rates Infrastructure as “Worst” on a MaxDiff task, that lowers its overall importance for the population, even if it might be very high among all the respondents for whom it is a job responsibility. Better would be to exclude it from choice tasks for her. In general, we conclude that respondents should be given an option to exclude items that are not relevant to their jobs. This is possible using the constructed list feature in Sawtooth Software for MaxDiff.

Additionally, respondents complain that unimportant items should be noted once, and subsequently, “Don’t bother me with that.” This aligns somewhat with recent findings on the MaxDiff method that prioritization is not unidimensional for *Best* and *Worst* items (Dyachenko et al., 2013) and that discrimination may be better in the *Best* direction. We conclude that it is advantageous to focus the tradeoffs on the subset of items that are closer to the Best end for any given respondent.

We considered using other variants of MaxDiff and concluded that they could solve some of the problems we encountered and not others. The negative respondent experience imposed by a lengthy survey could benefit from Express MaxDiff and Sparse MaxDiff (Wirth and Wolfrath, 2012). The unactionable results caused by too little differentiation of the top items could be addressed through Adaptive MaxDiff (Orme, 2006) or Bandit Adaptive MaxDiff (Fairchild et al., 2015). Independent of these methods, we’d still be left with the data quality problem.

CONSTRUCTION METHOD

To focus the survey items, we developed a method that builds on constructed list MaxDiff as follows:

1. Respondents are asked first to identify items that are irrelevant to their jobs (IRR).
2. Among the remaining, relevant items, they indicate items that are unimportant (UNI).
3. After removing irrelevant and unimportant items, respondents trade off importance among the important (or not-unimportant) and relevant items (REL).
4. To assist with scaling, data quality, and model identification, the MaxDiff item list REL includes a small number of items randomly selected from the IRR and UNI lists (RND).

This is an adaptive method within the survey, but not within the MaxDiff exercise (unlike the within-exercise approach described by Orme (2006)).

AUGMENTATION METHOD

Now, if we only use choices from the constructed list MaxDiff exercise, we would throw away knowledge; we also know implicitly what they disprefer. In steps 2 and 3 of the construction method above, we know that every item in REL is “better” than every item in UNI. Given that, we augment the data with choice tasks pairing each of the items from REL as “winning” over each of the items from UNI. Overall, we refer to this as “Constructed, Augmented MaxDiff” (CAMD).

Tasks that are irrelevant to their jobs (IRR in the construction method step #1) are not used to augment the data. However, because each of those has a chance of appearing among the RND items (step #4), we ensure that there will be some chance to control for inadvertent non-selection, and to ensure model identification in case none is selected as relevant.

We open-sourced our R code to create augmented tasks and analyze the data as part of the Rcbc package (Chapman et al., 2018).

RESULTS

At a high level, what we find is that the data appear to be of improved quality: we get 2 to 3.5 times as much choice data from the augmentation and respondents report a better experience.

We believe that the CAMD data are higher quality than the data we gathered with standard MaxDiff because respondents are giving more input on the items that they see as relevant to them. This is particularly advantageous when respondents have heterogeneous job responsibilities, as we see in our research on enterprise IT administrators and developers. We were surprised to see how severe the data quality problem can be. In one of our n=77 studies shown in Figure 1, 52% (n=40) respondents stated that at least half of the items were irrelevant to them!

Figure 1. Distribution of Percentage of Items that Respondent Marked as Relevant

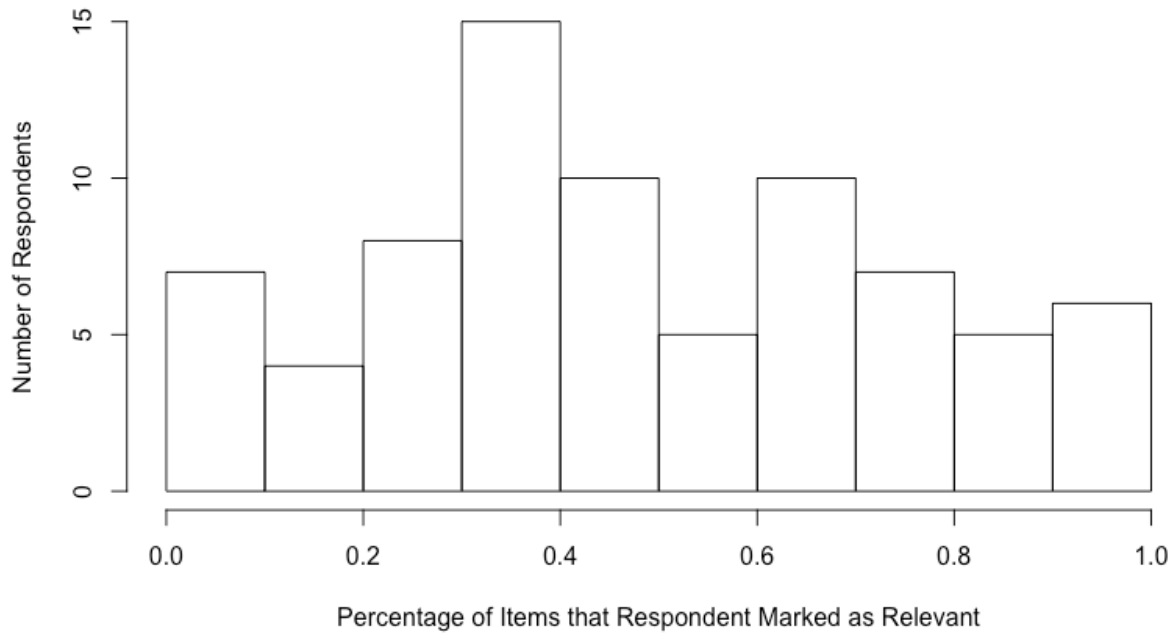
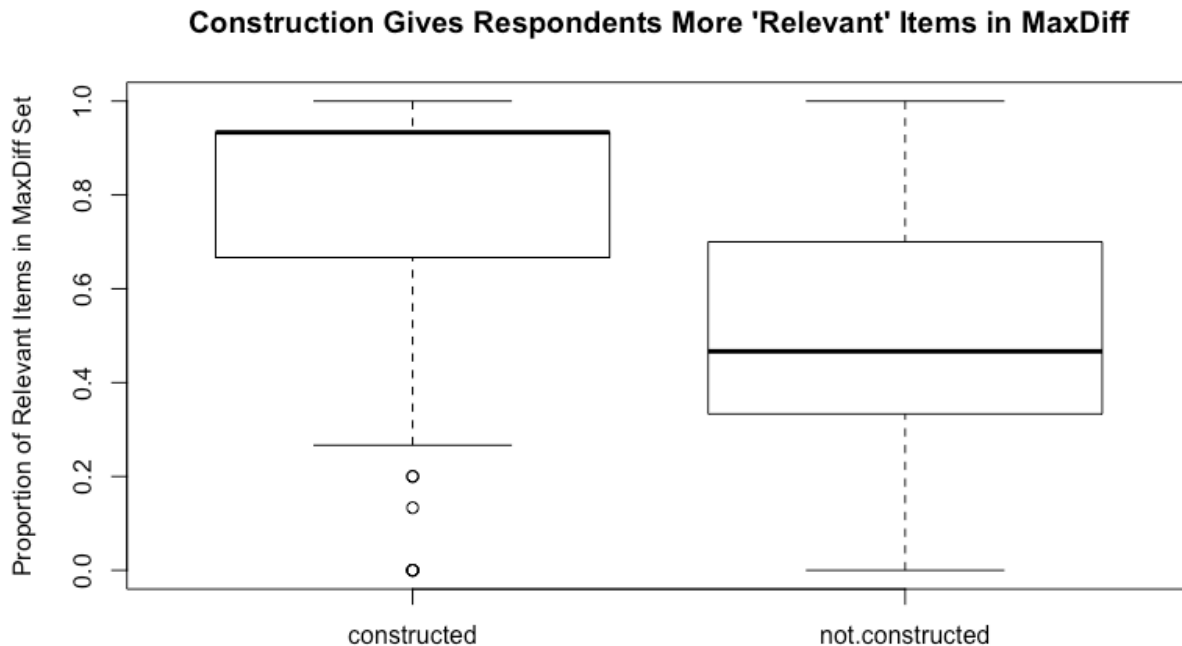


Figure 2 shows that constructing the MaxDiff item list reduced the amount of noise in our MaxDiff data by increasing the proportion of relevant items in the MaxDiff list from 46.7% to 93.3% for the median respondent. Without construction, the 15 MaxDiff items would be chosen randomly from the 30 candidate items.

Figure 2. Impact of Construction on the Proportion of Relevant Items in the MaxDiff Set for Each Respondent



Similarly, we found that construction increased the proportion of “important” items that were included in the MaxDiff exercise. In the same study as above, 66% (51/77) of respondents indicated that less than half of the items in the set were at least somewhat important to them. The distribution is shown in Figure 3. Construction doubled the proportion of important items in the MaxDiff exercise for the median respondent, from 33.3% to 66.7%, as shown in Figure 4.

Figure 3. Distribution of Percentage of Items that Respondents Marked as Important

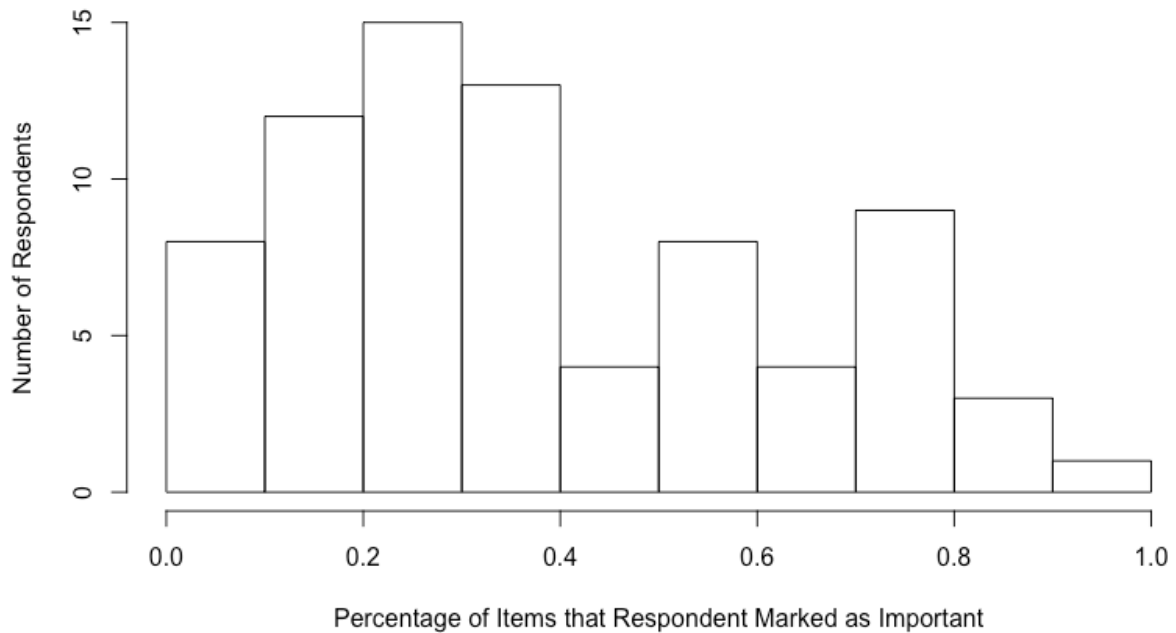
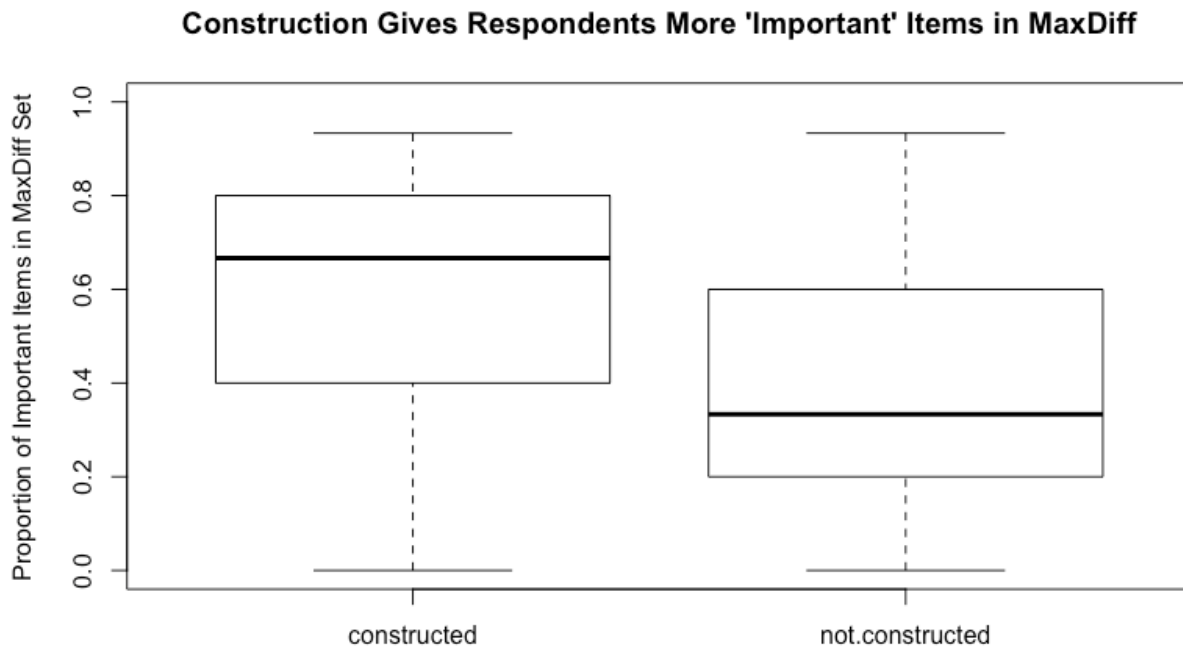


Figure 4. Impact of Construction on the Proportion of Relevant Items in the MaxDiff Set for Each Respondent



Respondent comments were almost universally positive for the CAMD surveys, including:

- “Great new setup.”
- “Seems faster this time.”
- “Thanks so much for implementing the ‘is this important to you’ section! Awesome stuff!”
- “I liked that this time around it was a lot quicker.”
- “Felt like the structure of the survey really locked in on my priorities much faster than previous surveys!”
- “Really pleased with the importance selections leading into the general survey. Big improvement IMO.”

The augmentation process led to modest adjustments in overall item utility scores (overall Pearson’s $r=0.82$ and 0.90 in two studies). Figures 5, 6, and 7 show the overall utility scores without and with augmentation for the study with $r=0.82$.

Figure 5. Overall Utility Scores without Augmentation

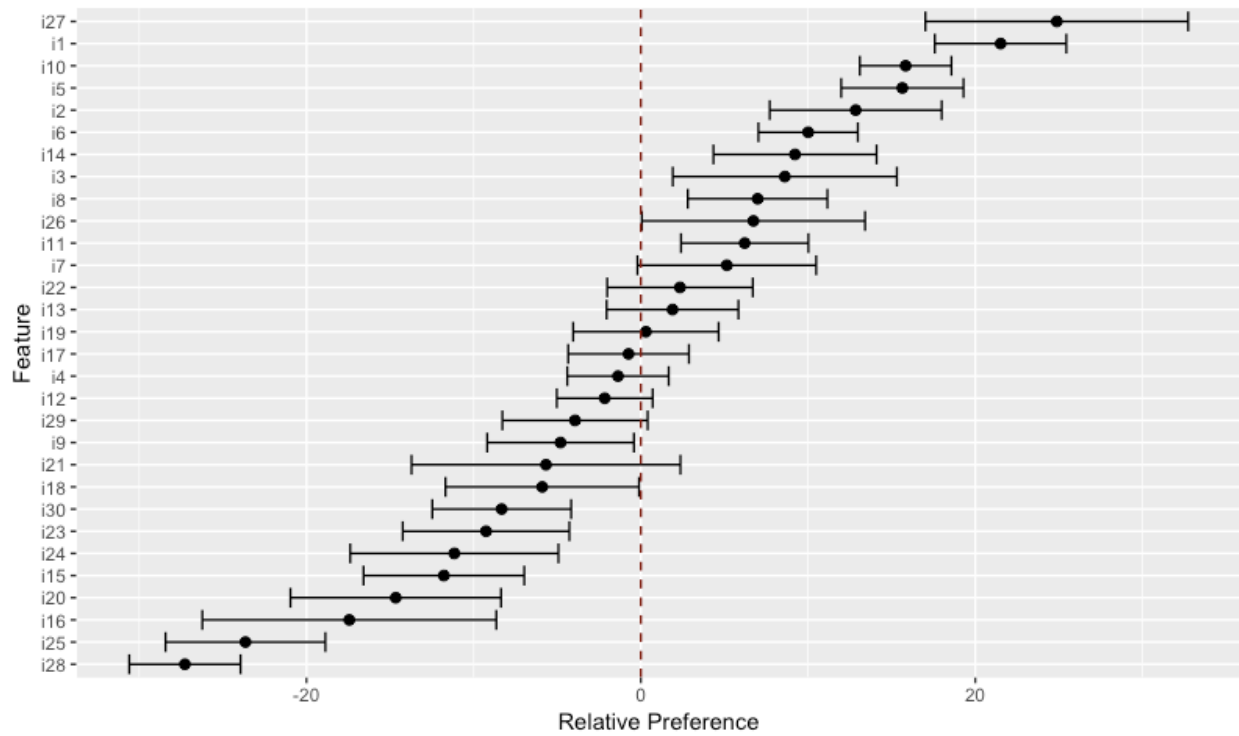
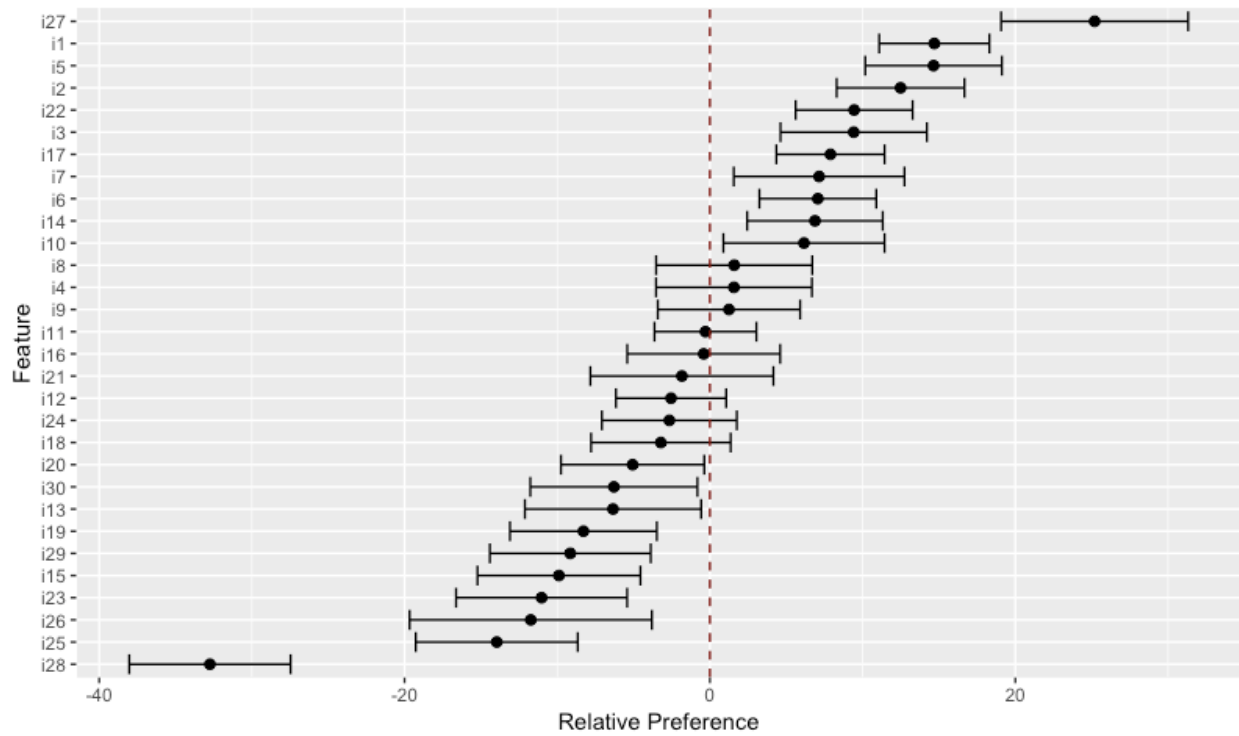
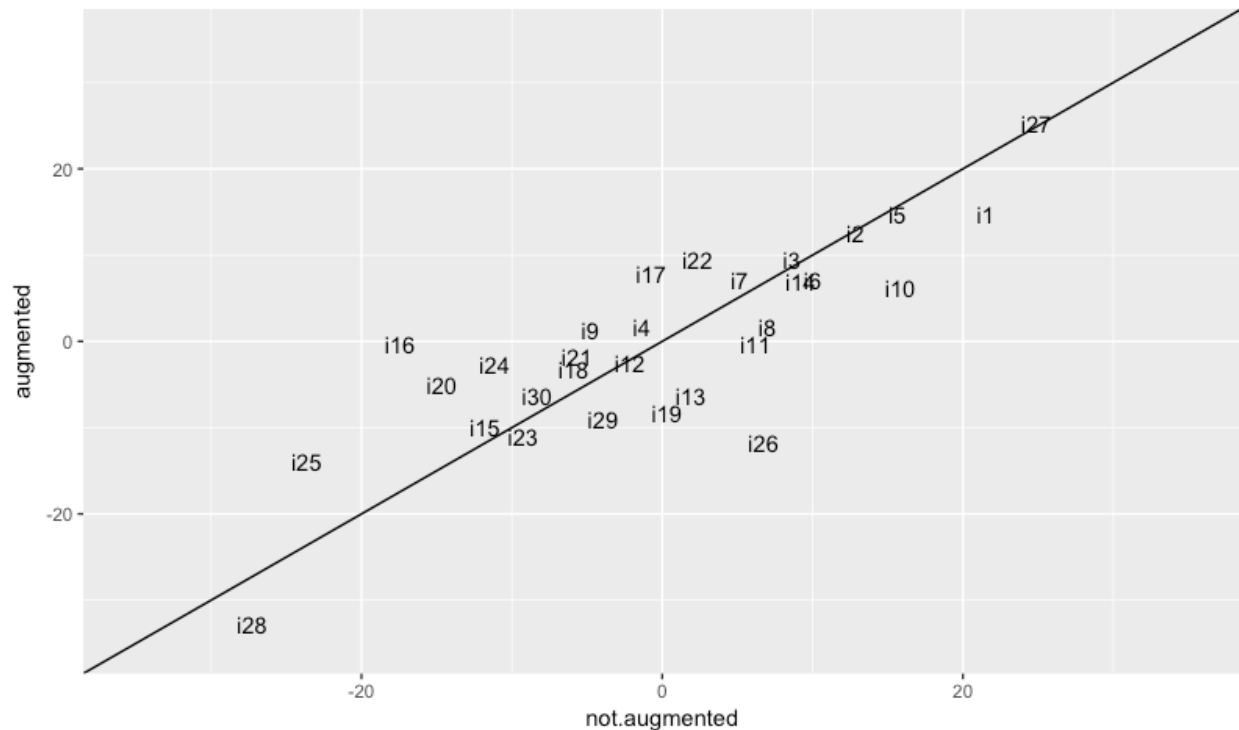


Figure 6. Overall Utility Scores with Augmentation



**Figure 7. Impact of Augmentation on Overall Utility Scores (Pearson's $r = 0.82$)
Overall (Upper-Level) Utility Scores without and with Augmentation for 30 Items**



Augmentation had more noticeable effects on individual-level utility scores, as expected. We examined the impact of augmentation on the individual scores of items for which the smallest and largest percentages of respondents reported them as unimportant. One item (“i3”) was marked as unimportant by 2.6% of respondents and two items (“i23” and “i28”) were marked as unimportant by 15.6% of respondents. For “i3,” Figure 8 shows that augmentation decreased the utility scores for individuals who marked the item as important and had high scores without augmentation. Augmentation had the opposite effect on “i3” scores at the lower end, namely increasing them. For “i23,” Figure 9 shows that augmentation mostly increased scores for respondents who marked it as important and mostly decreased scores for respondents who marked it as unimportant. There were exceptions for each category, though. For “i28,” augmentation had a more significant effect, decreasing scores of all respondents who marked the item as unimportant and raising most of the scores of respondents who marked it as important.

Figure 8. Impact of Augmentation on Individual Utility Scores for an Item (“i3”) Where 2.6% of Respondents Marked It as Unimportant.
Individual Utility Scores for One Item (i3) without and with Augmentation ($r=0.68$)

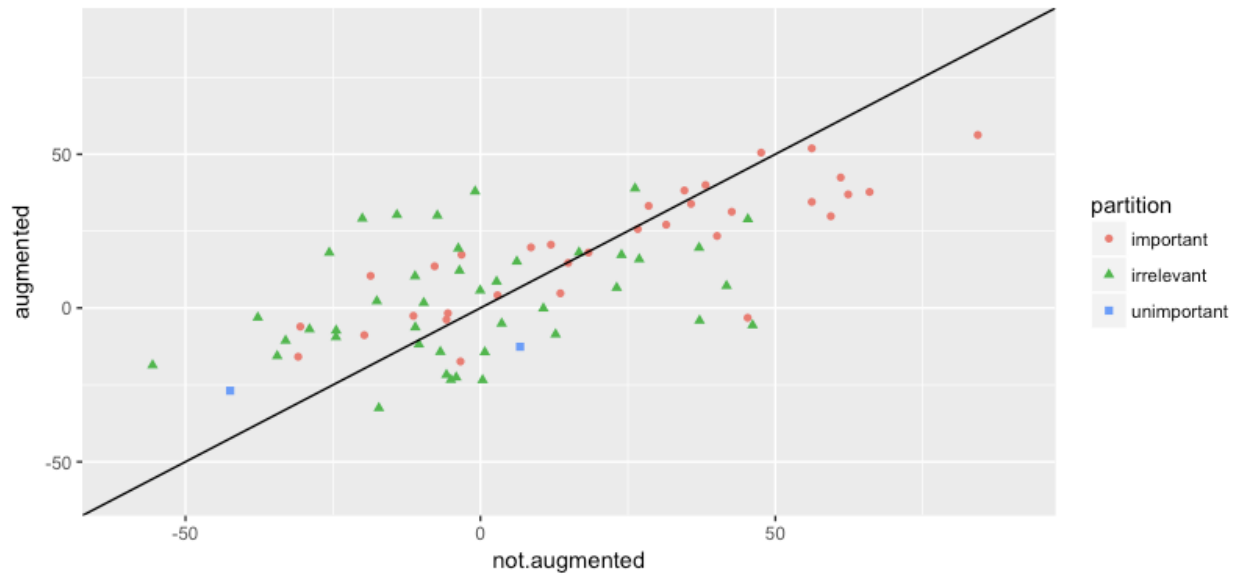


Figure 9. Impact of Augmentation on Individual Utility Scores for an Item (“i23”) Where 15.6% of Respondents Marked It as Unimportant.
Individual Utility Scores for One Item (i23) without and with Augmentation ($r=0.67$)

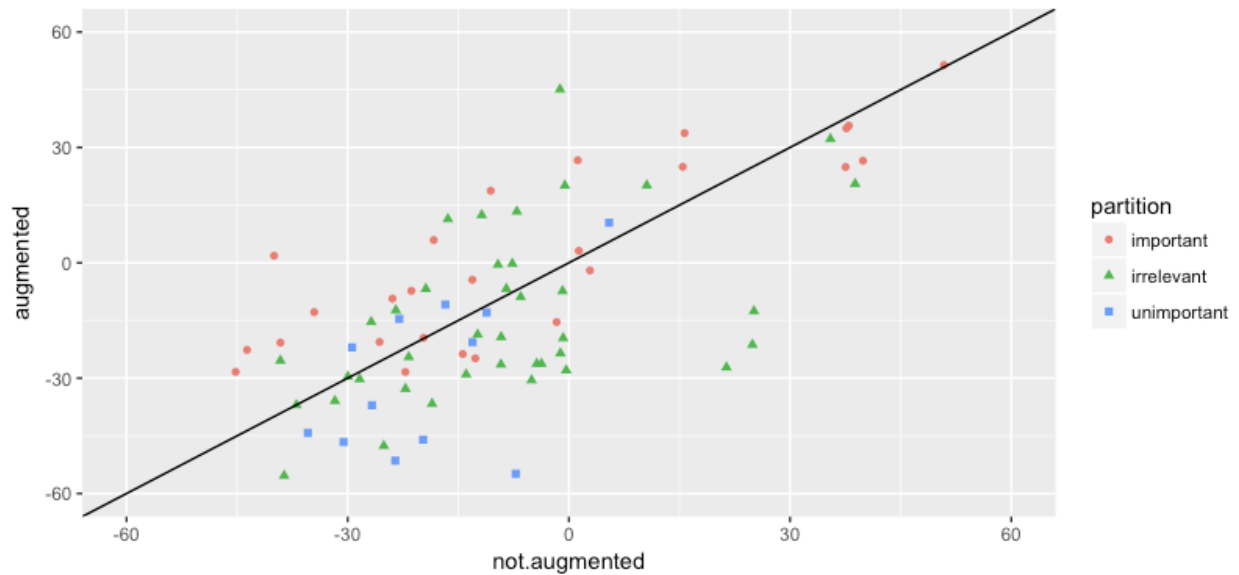
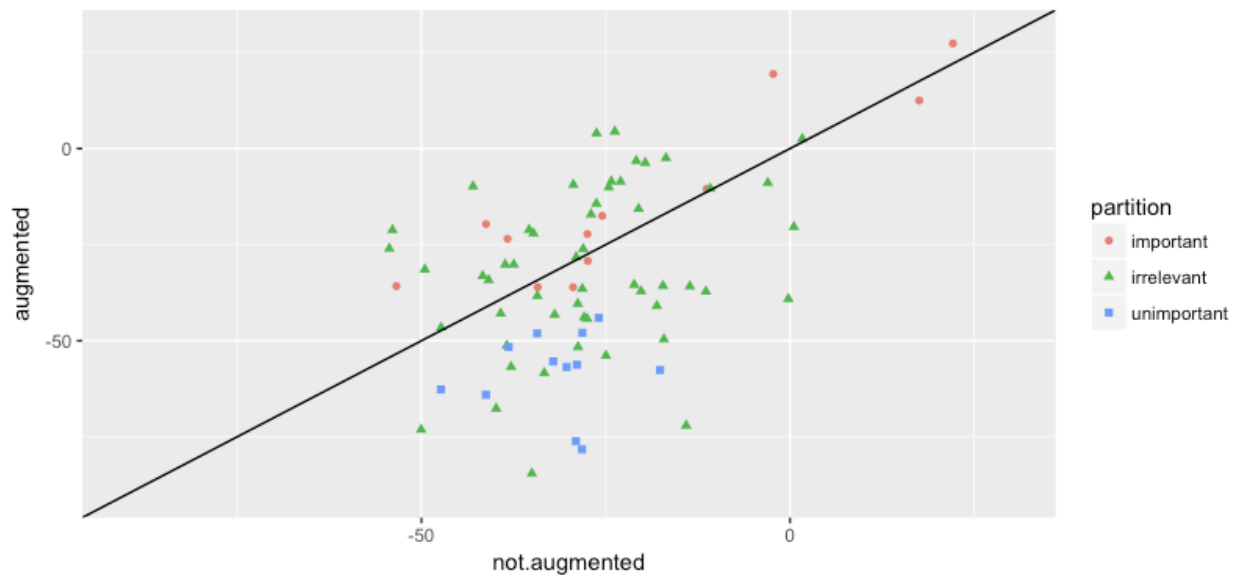


Figure 10: Impact of Augmentation on Individual Utility Scores for Another Item (“i28”) Where 15.6% of Respondents Marked It as Unimportant. Individual Utility Scores for One Item (i28) without and with Augmentation ($r=0.47$)



DISCUSSION

Relationship between “Importance” and MaxDiff Responses

We augment the explicit best/worst choices of respondents by creating implicit choices based on their responses to the “Importance” question. This uses all of the information that we receive from respondents, but also can put a lot of weight on their responses to the “Importance” question. If their responses to the “Importance” question are not consistent with their responses to the MaxDiff questions, then augmentation would amplify the inconsistency.

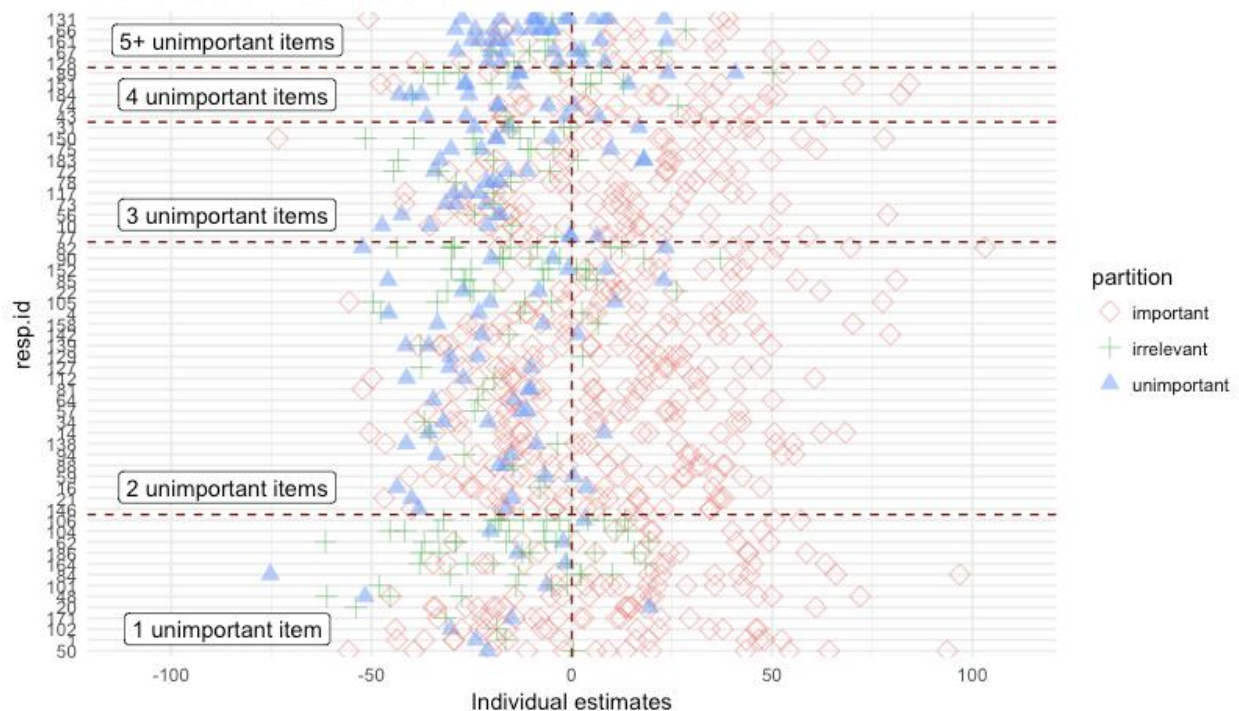
To what extent are responses to the “Importance” question consistent with the responses to the MaxDiff questions? We examined this in one study by comparing responses to the “Importance” question to the individual-level estimates from a hierarchical Bayesian model *without* augmentation. In this study, we presented a list of 30 items, of which 15 were included in the MaxDiff exercise per the “Construction Method” section above. Figure 11 shows the individual-level estimates for the 15 items that each of $n=77$ respondents saw in a MaxDiff exercise.

Individual-level estimates for UNI items¹ tended to be lower than individual-level estimates for REL and IRR items. To explore the patterns, we grouped respondents in the plot by the number of items that they marked as unimportant (1, 2, 3, 4, 5+). Within each such group we sorted by number of irrelevant items. Within the groups with 1-4 unimportant items, it appears that the individual-estimates of unimportant items are higher as the number of irrelevant items increases. More formally describing these patterns is an area for further research. It could be

¹ Our goal was to include two UNI items for each respondent, but some respondents had a different number. Those who had fewer than two were respondents who marked fewer than two items as UNI. Those who had more than two were respondents who marked fewer than $15 - 2 - 1 = 12$ items as REL (relevant and at least somewhat important).

insightful to examine what proportion of augmented tasks align with the unaugmented individual-level estimates.

Figure 11: Individual-Level Estimates without Augmentation by Partition (Important, Irrelevant, Unimportant)



Design Recommendations

We recommend that the constructed list include some IRR and UNI items in order to reduce the chance of an intractable model. If we didn't include any IRR items and one item was irrelevant to our entire sample, then the model would be intractable because we'd have no observations about that item. If we didn't include any UNI items and one item was unimportant to our sample, then it would have an unbounded negative utility. We typically add one IRR item and two UNI items to MaxDiff sets of 15-30 items. In 5+ CAMD studies, we have yet to encounter an intractable model, but we don't know whether one and two are the optimal numbers, or if the numbers should be a function of other study parameters (e.g., number of items, sample size). One approach to investigate this would be to estimate the model using only MaxDiff tasks that excluded IRR and UNI items and compare the model's stability to that of the model using all MaxDiff tasks.

The relevance and importance questions can be lengthy for respondents to answer, so we recommend breaking it into chunks, either on the same screen (by repeating column headers every few rows) or by splitting it into multiple screens. We hypothesize that the latter would increase response rates compared to the former because it seems less intimidating to respondents.

Open Topics

The impact of augmentation on individual-level estimates could benefit from more research because we saw different patterns on different items, as shown in Figures 8, 9, and 10 in the Results section. One pattern suggested by the estimates in those figures is that the impact of augmentation is more pronounced when more respondents mark an item as irrelevant. Both “i23” (Figure 9) and “i28” (Figure 10) were marked “unimportant” by 15.6% of the sample. However, we saw a greater impact of augmentation on “i28,” which was irrelevant to 68.8% of the sample compared to 50.6% for “i23.” It’s not clear to us why augmentation decreased the individual-level scores for some respondents who marked the item as important, as seen most prominently in Figure 8. Looking at Figure 7, we hypothesize that augmentation compressed the (zero-centered) scores for the middle $n-2$ items because it drove down the score for the lowest item (“i28”).

Could we save respondents time by asking only about “importance,” use that for both construction and augmentation, and omit the “relevance” question? Perhaps the two questions share a common set of underlying factors. We believe that this decision depends on the respondents and the items in the exercise. When respondents are more specialized relative to the items (e.g., surveying enterprise IT professionals about management tasks), then we propose keeping both questions. In other domains (e.g., consumer products), the distinction may be less relevant.

One signal we examined is the proportion of respondents who marked an item as UNI (i.e., relevant and not important). We expect that this proportion would be near zero in cases where relevance and importance are not different to respondents. We’ve seen 5-10% of respondents mark a given item as UNI in most of our studies. One study had UNI rates of 10-40%, which suggests that those respondents viewed the “relevance” and “importance” questions as especially distinct.

The optimal wording of the “relevance” question is still an open topic. We have primarily used two variants, depending on the audience:

- When surveying users or decision-makers directly, we’ve asked them to indicate whether the item is “relevant to your responsibilities or expertise.”
- When surveying their representatives (e.g., asking sales or support teams about their customers), we’ve asked them to indicate whether “you have visibility into the importance of the item to your customer.”

Is it possible that the augmented tasks will overwhelm the MaxDiff responses that respondents provide directly? If so, how could an analyst detect that and to what extent should augmentation be performed in such cases? The individual-level estimates of UNI items that were included in the MaxDiff set discussed above show that UNI items tend to have lower-estimates than REL and IRR items, but we don’t see that the lowest-rated item for each respondent is a UNI item. The impact of augmentation is related to the number of augmented tasks, which is a function of the ratio of REL:UNI items (the closer to 1.0, the more augmented tasks) and the proportion of items marked as IRR (the closer to 0.0, the more augmented tasks). These are, in turn, affected by a respondent’s likelihood to endorse an item.

CONCLUSION

The construction in Constructed, Augmented MaxDiff method increased the quality of our data by selecting more relevant items for each respondent's MaxDiff exercise. Additionally, respondents reported a more positive experience because they didn't spend so much time on items that were irrelevant or unimportant to them. Augmentation provided 2-3 times more choice data for our models without requiring respondents to make proportionally more trade-offs. We hope that our open-sourced R code makes it easier for others to apply the method to their domains.

ACKNOWLEDGEMENTS

We appreciate the improvements and clarifications suggested by Keith Chrzan at Sawtooth Software, who served as the reviewer for this paper. We're also grateful to Ula Jones for her thoughtful questions as this paper's discussant, such as asking about the need to include both the relevance and importance questions. Thanks also to our respondents, whose critical feedback motivated us to explore this method.



Eric Bahna



Chris Chapman

REFERENCES

- Chapman, C.N., Bahna, E., Alford, J.L., and Ellis, S. (2018). Rcbc: Marketing research tools for choice-based conjoint analysis and maxdiff, version 0.30. [R package] <http://goo.gl/oK78kw>
- Dyachenko, T., Naylor, R. W., and Allenby, G (2013). "The Ballad of Best and Worst." In Orme, B., ed. (2013), Proceedings of the 2013 Sawtooth Software Conference, 357-366. Available at <https://www.sawtoothsoftware.com/download/techpap/2013Proceedings.pdf>
- Fairchild, K., Orme, B., and Schwartz, E (2015). "Bandit Adaptive MaxDiff Designs for Huge Number of Items." In Orme, B., ed. (2015), Proceedings of the 2015 Sawtooth Software Conference, 105117. Available at <https://www.sawtoothsoftware.com/download/techpap/2015Proceedings.pdf>
- Orme, B. (2006). "Adaptive Maximum Difference Scaling." Technical paper series, Sawtooth Software. Available at <https://www.sawtoothsoftware.com/support/technical-papers/maxdiff-best-worst-scaling/adaptive-maximum-difference-scaling-2006>
- Wirth, R. and Wolfrath, A. (2012). "Using MaxDiff for Evaluating Very Large Sets of Items" In Orme, B., ed. (2012), Proceedings of the 2012 Sawtooth Software Conference, 59-78. Available at <https://www.sawtoothsoftware.com/download/techpap/2012Proceedings.pdf>

SHAPLEY VALUES: EASY, USEFUL AND INTUITIVE

DAVID W. LYON

AURORA MARKET MODELING, LLC

INTRODUCTION

Shapley Values¹ (SVs) are a general and widely useful way to summarize how much individual items contribute to the overall value of combinations of items (e.g., product assortments, feature bundles, sets of advertising claims). Although they have been discussed and promoted for marketing research use since at least 2000 (Conklin and Lipovetsky, 2000, 2005, 2013; Conklin and Shmulyian, 2012), SVs appear to be used far less often than they could and should be. This paper seeks to build an intuitive understanding of them (as opposed to relying on mathematical formalisms) to encourage their wider use and to address some computational issues, including a too-little-known trick for fast and exact computation in many cases.

In marketing research applications, Shapley Values are not a standalone analytic technique. They assume some other agreed-upon way of evaluating the value of a combination of items and build on that analysis by summarizing the effects of individual items. One common example is TURF analysis, where the value of a set of items is its combined unduplicated reach. Another is key driver analysis, where the total regression *r*-squared produced by a set of predictors is the value of that set of items. While both of these are common applications, Shapley Values are in no way restricted to those two. They are useful in almost *any* analysis based on the value of combinations, no matter how that value is defined or determined.

The fundamental usefulness of Shapley Values is that they shift the focus from *combinations* of items, of which there are often billions or more, to the *items themselves*, which are few enough for a human being to deal with and think about. Sometimes, as in the literal TURF problem of finding the one best combination of a given size, this is irrelevant. More often, however, it is crucial to both analysts and business managers as a way of summarizing and understanding what is going on with items. The Shapley Values can be thought of as providing an overview or “road map” to TURF or other combination-based analysis.

We will deliver on the title promises of “easy, useful and intuitive” in reverse order. The first section addresses the intuition, the second discusses usefulness (a relatively obvious point once we have an intuitive understanding), and the third looks at computational issues and how to make them easier. A final section looks at the special topic of TURF on MaxDiff data, whether and when Shapley Values help there, and what their behavior there implies about the usefulness of TURF on such data.

¹ In the academic game theory literature, the “Shapley Value” (singular) is actually a *set* of values, one for each player in a game. With that acknowledged, we will use the typical marketing research jargon where a “Shapley Value” is for just one item of interest, and the whole collection of them are “Shapley Values” with a plural *s*.

INTUITION: WHAT ARE SHAPLEY VALUES?

Let's begin by considering key driver regressions (KDRs), which are widely used in customer satisfaction work, among other areas. They seek to predict some overall measure, often overall customer satisfaction, based on ratings of a number of individual items, often satisfaction with particular aspects of a product. Beyond the simple prediction, the goal is to determine which items are most important and to quantify the importance of each.

One obvious idea is to enter the individual items into a regression sequentially and observe how much the overall r-squared, or variance explained, increases as each is added. That increase can be viewed as the contribution, or importance, of that item. But, the item ratings are not statistically independent—they often are strongly collinear—so the order in which they are entered into a regression has a huge effect on the apparent importances. In effect, the first item entered gets credit for all the shared variance that might equally well have been explained by others, while the last entered gets credit only for its own unique contribution independently of the others. If the order makes such a big difference, what is the right order?

There simply is no “right” order. An elegant answer to this issue is to consider all possible orders, and average the item importances over all the possibilities. This idea was introduced by William Kruskal (1987) and is known today as “Kruskal relative importance” or “Shapley Value regression,” among other terms. It is appealing in that it treats all items identically and fairly, with each being first equally often, as well as second, third, . . . , and last equally often. There is no right order, so we average the order out of the question.

The Shapley Value for an item in this situation is simply the average amount by which it increases the regression r-squared, averaged over all possible orderings. This is illustrated in Exhibit 1 for a very small example.

**Exhibit 1. Example of Key Driver Regression with 3 Items—
Averages Are Shapley Values**

Incremental r^2 by order of entry												Averages (SVs) ²	
1: A-B-C		2: A-C-B		3: B-A-C		4: B-C-A		5: C-A-B		6: C-B-A		A	0.16
A	0.26	A	0.26	B	0.22	B	0.22	C	0.17	C	0.17		
B	0.09	C	0.06	A	0.13	C	0.06	A	0.15	B	0.11	B	0.12
C	0.01	B	0.04	C	0.01	A	0.08	B	0.04	A	0.08	C	0.08

Our interest here is in using key driver regressions to motivate the averaging-over-orderings interpretation of Shapley Values, not in the KDRs themselves. While Shapley regression is one reasonable way to perform KDRs, it is by no means the only or the most efficient. Readers interested in KDRs *per se* should see Cortese 2018 (in this volume) or consider random forests or any of many other approaches to KDRs. “Relative weight analysis” (Johnson and Lebreton, 2004) produces results remarkably similar to Shapley

² Many practitioners would re-express the Shapley Values—the average r-squared contributions—as a percentage of the overall r-squared. While popular, that mostly serves to obscure their natural interpretation. There is no inherent reason that any importance measure should total 100%.

regression, but with very different math and far less computational burden, making larger problems feasible.

ORDERINGS VS. COMBINATIONS

Our introductory claims about Shapley Values were about *combinations* of items, while the KDR example revolves around *orderings* of items. Why the discrepancy? Despite the naturalness (and frequent usefulness) of an averaging-over-orderings view of Shapley Values, they are fundamentally about combinations rather than orderings.

Consider a concrete example of 13 items in the order A-M-K-D-F-B-L-E-C-J-H-I-G. Focus for the moment on item “F.” F’s contribution to r-squared in this ordering is entirely independent of the ordering of the 8 items behind it (B-L-E-C-J-H-I-G). Those could be in any of 8! (“8 factorial” or 40,320) different orderings, in all of which F adds the identical amount to r-squared. For that matter, the amount F adds to A-M-K-D depends heavily on those being the particular four items entered in the regression ahead of it, but not at all on the *order* in which those four were entered. They could be in any of $4! = 24$ different orders, with no effect on the contribution of F. In effect, we are just looking at how much better the combination {A,D,K,M,F} is than the combination {A,D,K,M}, and remembering that the improvement F offers there will apply in $24 \times 40,320 = 967,680$ different orderings.

From this point of view, Shapley Values are averages over combinations, with appropriate weights to reflect how often each combination turns up in all orderings. Specifically, the Shapley Value for item F is the *weighted* average, over all combinations that don’t include F, of the r-squared gain when F is added to those combinations. If there are n items, and k of them are ahead of F (i.e., we are adding F to a combination of size k), the weight is $k! (n - k - 1)!$, since there are $k!$ orders for the k items ahead of F and $(n - k - 1)!$ orders for the $(n - k - 1)$ items behind it.

The full formula for Shapley Values from this point of view is the one most often seen in the marketing research literature. It looks like

$$\varphi_i(v) = \sum_{S \subseteq (N-i)} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)]$$

Here, $\varphi_i(v)$ denotes the Shapley Value for item i relative to a “value function” v (for KDRs, the r-squared of a regression). N is the set of all items being considered and $|N|$ is its size (n in the earlier notation, the number of items). S is some subset of N that does not include i , and $|S|$ its size (k in the earlier notation).

The key term is $v(S \cup \{i\}) - v(S)$, the value of (i.e., r-squared for) set S with i added to it, minus the value of S alone. The term before it is the weight, divided by the sum of all weights (which turns out to be $|N|!$, the total number of orderings of N).

WEIGHTS AND COMBINATION SIZES

It is easy to overlook or misinterpret the importance of the weighting by number of combinations. One oversimplification is to say that the Shapley Value is the average contribution of an item to a combination it is not already part of, averaged over all

combinations. This ignores the weights and puts a strong emphasis on mid-sized combinations because there are so many of them, but virtually none on very small or very large combinations, which are few in number.

The opposite mistake has also appeared in the literature, where the huge weights of $(n - 1)!$ that apply to the largest and smallest combinations are taken to swamp the small weights for mid-size combinations and imply that Shapley Values can be closely approximated by averaging the contributions to the two extreme-size combinations.

In fact, the number of combinations of each size, times the weight applying to each combination of that size, is constant across sizes! In effect, the Shapley Value is the average contribution of an item to combinations it is not already part of, with all sizes of combinations weighted equally. Exhibit 2 gives a small concrete illustration.

Exhibit 2. Number of Combinations and Their Weights for $n = 9$ items

Size of a combination not including a particular item k	Number of such combinations of size k $\binom{n-1}{k}$	Weight on each $k! (n - k - 1)!$	Total weight for size k (number \times weight)
0	1	40,320	40,320
1	8	5,040	40,320
2	28	1,440	40,320
3	56	720	40,320
4	70	576	40,320
5	56	720	40,320
6	28	1,440	40,320
7	8	5,040	40,320
8	1	40,320	40,320

We can emphasize the equal weighting of combination sizes by rewriting the formula for a Shapley Value as:

$$\varphi_i(v) = \frac{1}{|N|} \sum_{k=0}^{|N|-1} \left(\frac{1}{\binom{|N|-1}{k}} \sum_{S \subseteq (N-i), |S|=k} [v(S \cup \{i\}) - v(S)] \right)$$

The inner sum here is over all subsets S of size k that don't include i . There are no weights involved; the term in front is simply dividing the sum by the total number of such combinations. The inner sum and its divisor represent the average contribution of item i to combinations of size k (not counting i). The outer sum is over all combination sizes; there are $|N|$ sizes, ranging from 0 (i.e., item i is added to the null set) to $|N| - 1$.

In this formulation, the Shapley Value is an unweighted average of unweighted averages, an appealingly simple way to look at things.

THE INTUITIONS

We have seen three ways of expressing or thinking about Shapley Values. All are mathematically equivalent, but differ in their usefulness.

First, they are the average contribution of an item, as items are added sequentially, over all possible sequential orders.

Second, they are the contribution of an item when added to a combination of items that doesn't already include it, averaged over all such combinations with appropriate weights.

Third, they are the contribution of an item when added to a combination of items that doesn't already include it, with all combination sizes counted equally.

The first (orderings) interpretation is natural in the context of key driver regressions and some other situations, and often helpful in reasoning about Shapley Values. But it is computationally unfriendly in that there are far more orderings than combinations (e.g., for 20 items, 2.4 quintillion orderings, but only about a million combinations).

The second version is the most commonly seen formulation, but perhaps the least helpful intuitively. It is crucial to understand the role of the weights it employs.

The third interpretation is the most natural in many applications, including TURF. We will see that in some cases, it is a helpful computational framework as well. Like the first, it can be stated precisely in simple words.

SHAPLEY VALUES AND TURF

TURF may be the most common use context for Shapley Values in marketing research, and involves some interesting issues. While TURF stands for Total Unduplicated Reach and Frequency, practice almost always focuses on the unduplicated reach and ignores the frequency.³ The unduplicated reach of some combination of items in a pure TURF formulation is simply the percentage of respondents for whom at least one of the items “reaches” the respondent or “is a hit.”

If the items are flavors or varieties of a product, an item “reaching” a respondent might mean that the respondent is willing to buy that flavor or variety. If items are messages or ad claims, an item might be considered “a hit” if the respondent would believe that claim or considers it important. If items are possible product features, being a hit might mean that the feature is wanted by the respondent. There are endless variations, but in all cases the respondent data is 0 or 1 for each item for each respondent: either a respondent would buy, believe, want, etc. the item or she would not.

A *combination* of items is considered to reach a respondent if there is at least one item the respondent would buy, or at least one claim she would believe, or at least one feature he wants. This is a grossly simplistic formulation of many real marketing problems, but TURF is nonetheless in wide use.

A slight modification of TURF is to vary the “depth,” or the number of items in a combination that must reach a respondent for the combination to be considered a reach or

³ In many situations, frequency questions are not even asked.

success. With a depth of 3, say, at least three items must be hits for the combination to count as a reach. Increasing the depth tends to reward items that co-occur with others to achieve the required depth, rather than those that may have little overlap but reach entirely new respondents. This is especially true for smaller combinations.

The canonical goal of a TURF analysis is to find “the” best combination of a given size—the one with the highest reach. Typically, the top 10 or top 100 or so combinations are found and presented for each size of interest.

TURF is a challenging problem because the best combination of k items is not necessarily just the k with the highest individual reaches (except for $k = 1$). Similarly, the best combination of k is not necessarily the best combination of $k - 1$ plus one more item. (Assuming otherwise leads to a “stepwise TURF” analysis with no guarantee of obtaining the true TURF answer.) If two items overlap considerably in which respondents they appeal to, high-reach combinations are likely to include only one of the two. The overlap patterns are fundamental to the results.

Consequently, the only straightforward way to find the best combination of k is to enumerate and evaluate all the possibilities. This is easy if n and k are small, but becomes computationally difficult or impossible as they increase. In practice, n is usually restricted to about 25, or perhaps 30 at the most, for a full analysis.

Typical TURF results might look something like Exhibit 3, which shows an analysis for 13 items and the best combinations of size 2 to 5 only. With 20 or more items and combination sizes up to 15 or 20 or more, potentially repeated for several different depths, the size and number of tables like this rapidly balloon.

Exhibit 3. Example of Typical TURF Results

Sample of 300 physicians surveyed re pharma company communication channels
Ten best combinations and ties for each size of combination

Pairs $k=2$			Triples $k=3$			Quads $k=4$			Quints $k=5$		
Rank	Reach	Items	Rank	Reach	Items	Rank	Reach	Items	Rank	Reach	Items
1	77.1	GD	1	81.7	GDM	1	83.4	GBMH	1	84.2	GBCMH
2	75.3	GB	2	81.6	GBM	1	83.4	GDCM	1	84.2	GDCMH
3	74.4	BM	3	81.4	GBD	3	83.3	GBDM	1	84.2	GBDCK
4	74.3	DC	3	81.4	GDC	3	83.3	GBMK	1	84.2	GBDCH
5	73.6	BD	5	80.4	GBK	5	83.1	GBDK	5	84.0	GBDCM
6	72.5	GM	6	79.9	GCM	5	83.1	GBDC	5	84.0	GBDMH
7	70.8	BK	7	79.9	GBL	7	82.8	GDCH	5	84.0	GBCMK
8	70.7	GC	8	79.8	GDK	7	82.8	GDCE	5	84.0	GBMKH
9	69.4	BL	9	79.7	GDL	7	82.8	GDMJ	5	84.0	GBMLH
10	69.1	BC	10	79.5	GDE	10	82.7	GBDL	10	83.8	GBDCL
						10	82.7	GBCM	10	83.8	GBDKH
						10	82.7	GBDH			
						10	82.7	GDMK			

The top line of this table answers the literal, nominal, TURF question of which one combination is best. This is seldom the real, or only, objective, however, which is why tables of multiple top combinations are typically produced. The problem with these tables is that it is difficult to see or extract general patterns, conclusions or insights beyond the literal reading of “GDM is best, GBM is second best, GBD and GDC are tied for third.” Such readings are seldom managerially useful and certainly not interesting.

Let’s see what Shapley Values could contribute here. We can compute them using the same formulas presented for key driver regressions. The key difference is that the “value function” v in those formulas will now be the TURF value of a combination—the percentage of respondents for which a combination is a reach—rather than the r-squared from a KDR.

Another difference is that we can define that value function at the individual respondent level, as 1 or 0, depending on whether a particular respondent is or is not reached (at the specified depth, if applicable). The aggregate value function is then just the average over the individuals.⁴ This is a convenient way to formulate things: it makes it trivial to apply respondent weights, if desired, and to compute (weighted) standard errors for the reach percentages. It is also critical to a computational approach we discuss in a later section.

⁴ For KDRs, of course, there is no respondent-level analog to the r-squared and this is not possible.

Exhibit 4 shows the Shapley Values for the same TURF data used for Exhibit 3. In a single table we see results for three different depths of TURF at once, and it is a full analysis based on all combinations and all items. There is even room left for meaningful item labels.

Exhibit 4. Shapley Values for the TURF Results of Exhibit 3

Standard	TURF	TURF	Information Source	
13.8	11.8	11.4	G	In-person sales rep visit
13.8	13.2	12.8	B	Informative snail mail
11.7	11.7	10.7	D	Emails, E-newsletters
11.6	10.8	10.8	C	Leaflet/brochure in snail mail
8.0	8.0	8.0	A	Website
7.1	6.9	6.4	M	Smartphone apps
4.8	4.4	4.4	K	Self-guided online programs
4.4	4.4	4.2	L	Reminder/alert/pop-up on <xxx>
4.3	3.9	3.4	H	Live talk with tech experts
1.9	1.9	1.9	E	Texts from company
1.4	1.4	1.2	J	Live-assisted online programs
1.4	1.4	1.4	F	Phone call from sales rep
0.9	0.9	0.9	I	Robo calls

From the table of Shapley results, it is easy to see that G and B are the two items that contribute the most, that they are close to each other in contribution and that B is a bit better at the higher depths. Similarly, we can see that D and C occupy a second tier, also very similar to each other in effect, and the other items taper off from there, with items E, J, F and I at the bottom contributing little.

USEFUL: WHAT'S THE BENEFIT?

Dealing with results for 13 items, rather than the best combinations out of 8,192 possible ones, with different depths laid out in parallel and items labeled, makes tables like Exhibit 4 far more informative and makes it possible to spot general patterns. Shapley Values are easier to think about, easier to present and more memorable than endless lists of combinations. That, in a nutshell, is the power and usefulness of Shapley Values.

Note that the Shapley Value analysis does *not* answer the actual TURF question of which combination is best. The *k* items with the highest Shapley Values do not necessarily make up the best combination of *k*.⁵ If the classic TURF question of “what’s best?” is in fact the key objective, Shapley Values don’t address it. But if the goal is to *understand* item values in the context of combinations, rather than to find the one best

⁵ Their chances of doing so are better than the chances of the *k* items with the highest individual reach, or the *k* items from a stepwise TURF analysis, but there is no guarantee; a comprehensive search is still needed to find the definite best combination.

answer, the Shapley Values are far easier to work with. While clients often ask for the one best combination, their true underlying objective is likely to be an understanding or insight into item values or patterns.

STABILITY OF RESULTS

One appealing aspect of Shapley Values is that they are far more stable than analyses relying on individual combinations or orders. In TURF analysis, changing a single respondent's answers can cause upheaval in the rankings of combinations, and certainly in which one is best. This is particularly true for larger combinations whose reach approaches the maximum possible for the sample—they tend to “chase” individual respondents by including items with unique if limited appeal. In KDRs, approaches based on stepwise regression (i.e., on a particular ordering) show similar instability.

Those instabilities mean that TURF or stepwise regression, as well as many other analytic approaches to combinations, can't be expected to produce similar results across waves of a tracking study, or across subgroups of respondents, even if there is in fact no underlying change or difference. This is a problem common to many forms of ranking-based analyses. Shapley Values, however, are averages of interval-scaled quantities, not ranks. They are far more stable in the face of minor data changes, meaning that changes in results are far more likely to be real. Further, their standard errors can be computed in some situations (discussed below), facilitating formal significance testing and confidence intervals.

OTHER BENEFITS? NO.

Other claims of benefits are sometimes made for Shapley Values. There are two in particular that this author investigated and intended to present, but found that they did not hold up.

One is the idea that using combinations of product flavors, say, based on the top Shapley Values will result in better performance in the face of real-world out-of-stock situations. Suppose in an ice cream market that the huge majority of people like vanilla and many of those also like chocolate and/or strawberry. But say a few people like only mango, and a few others only red bean ice cream. A straight TURF solution might turn up a vanilla-mango-red bean combination as best. But if vanilla goes out of stock, the majority of the market has no acceptable choice left. The top Shapley Values would likely go to vanilla, chocolate and strawberry and if that combination is stocked, the sales loss from an out-of-stock on vanilla would be far less.

This argument has intuitive appeal, but does not hold up consistently in empirical data. It is generally true that a combination of the top Shapley Values loses less of its reach when one item goes out of stock (on average, across all items in the combination) than does the exact TURF optimum (assuming they are different in the first place). But that is not good enough for the SV-based combination—it must lose so much less with one item gone that the smaller loss makes up for any initial underperformance. That is a high bar that often cannot be met.

One simple way to move in the direction of better out-of-stock performance is to use a depth of 2 in the TURF search for the best combination rather than the standard depth 1. While this is too conservative, and means that the fully-stocked performance may well be less than optimal, it does help with out-of-stock performance.

An even better, straightforward, and precisely targeted approach is to redefine the value function used to evaluate combinations. Instead of unduplicated reach or depth=2 duplicated reach, we can define something we might call “resilience”: average reach with one item out-of-stock (averaging over all items in the combination). We might also consider a value function that is a weighted average of the pure unduplicated reach and the resilience. Then an exhaustive search of all combinations, in the same manner as a standard TURF search, will find the combination with the best out-of-stock performance. Indeed, tailoring the value function to the desired outcome is a flexible and important idea: the simplicity and possible convenience of a pure TURF value function should not dictate its blind use.

In any event, Shapley Values do not inherently help achieve the byproduct of out-of-stock resilience. Of course, Shapley Values can be computed on some sort of resilience as a value function, providing all their usual benefits of simple summarization.

In large TURF problems, it is not possible to evaluate all possible combinations. This leads to heuristic search procedures of various kinds: greedy searches, Federov swaps, genetic algorithms, etc. Another idea about Shapley Values is that the combination of the items with the top Shapley Values would be a good starting point for such searches. In this author’s experience that is true: it is an excellent starting point. Further, the Shapley Values can be used to guide the search further. While such an approach definitely outperforms starting from random combinations (it converges to a solution more rapidly, and more consistently to the true best), it does not seem to offer any consistent advantage over starting from a stepwise TURF solution. Stepwise solutions are trivial and fast to compute, even more so than Shapley Values, and sometimes far more so.

In sum, Shapley Values are useful because they condense and summarize the behavior of items in combinations. That is a major benefit; further beneficial side effects are not required to justify their use, although their stability is nice. If side benefits like resilience are desirable, they can be obtained directly by defining the appropriate criterion as the value function in a search over combinations, but they do not “magically” drop out of the Shapley Value computation.

EASY? COMPUTATIONAL ISSUES

There is a computational elephant in the room: computing Shapley Values requires evaluating all possible combinations of items. At least, that’s what the straightforward formulae assume. This is fine for 20 or so items, perhaps OK for 25 or so. But by about 30 items, a typical PC will have problems even generating and enumerating all the combinations, let alone evaluating them.

There are two ways around this issue. The more obvious is to work with a random sample of combinations rather than all of them. The second subsection of this section of the paper deals with a few details of doing that. The more interesting and novel way is a

trick that only works for some value functions, but offers super-fast, exact results even with huge numbers of items. TURF, including at varying depths, and many related value functions are among those where this approach works. We consider it first.

A FAST SHORT-CUT TO COMPUTING SHAPLEY VALUES

Let's return to the idea of TURF at the respondent level. (Surprisingly enough, the answer to our intractable aggregate computation problem does in fact lie in doing the computation for each individual!) Consider the formulation of SVs as the average over combination sizes of the average contribution to a combination of a given size.

If there are n items and a particular respondent has hits on h of them, how do the items contribute to combinations of size k ? To be concrete, let's say there are $n = 6$ items, $h = 2$ of them are hits and we are interested in combinations of size $k = 2$. What is the average contribution—the Shapley Value at the respondent level—of each item?

For the $n - h$ (4, here) non-hit items, the answer is simple. Adding them to a combination can't increase the reach, so their SVs are always zero.

But what about the h (2) hit items? Focus on what happens when a particular one of them is added to a combination of k (2) it is not already included in. That means the potential combinations of k (2) it could be added to will be all those composed of the remaining $n - 1$ (5) items, of which $h - 1$ (1) are the other hits. The item of focus will create reach where there was none before, generating a Shapley Value contribution of 1, when and only when neither of the items in the combination of k (2) is a hit. If any one of the k (2) is already a hit, the TURF value doesn't change, so there is no contribution to the SV. Thus, the Shapley Value contribution at this size of combination will be simply the probability that a random combination of k items, out of $n - 1$ total items, contains none of the $h - 1$ hits that are among the $n - 1$.

That probability is the *hypergeometric* probability of zero hits in a sample of k out of $n - 1$ that includes $h - 1$ hits, denoted and defined as ($x = 0$ denoting zero hits among the k):

$$p_{\text{hypergeometric}}(x | h - 1, n - 1, k) = \frac{\binom{h - 1}{x} \binom{n - 1 - (h - 1)}{k - x}}{\binom{n - 1}{k}}$$

The three combinatorials in this formula (each an “n pick k” evaluation) look a bit ugly on the page, but are trivial to compute.

The classic description of the hypergeometric is in terms of sampling without replacement from an urn containing, in this case, $n - 1$ balls, $h - 1$ of which are black (“hits”), leaving $(n - 1) - (h - 1) = n - h$ that are white. The calculated probability is that of obtaining x black balls in the sample of k . In our concrete illustration, it tells us the probability of zero hits in a sample of 2 items from a total of 5, 1 of which is a hit and 4 of which are not. Equivalently, it is the proportion of all combinations of size 2 that contain no hits.

Exhibit 5. Logic of the Respondent-Level Hypergeometric Calculation

Conceptually	Concretely	Visually
We have n items, of which h are hits, $n - h$ are not	Let's say $n = 6$ $h = 2$ hits	①②③④⑤⑥
Consider a single hit, leaving $n - 1$ others, of which $h - 1$ are hits.	Pull one hit aside. 5 items are left, 1 of them a hit.	①②③ ⑤⑥ ④
Consider combinations of size k	Let's say $k = 2$	
What happens when we add the selected hit to a combo of size k ?	What happens if ④ is the third one in?	? ? ④
If none of the k are hits, reach goes from 0 to 1, SV contribution is 1	In this case, our item "scores the win"	○ ○ ④
If another hit is already among the k , reach stays 1, no SV contribution	In this case, we've been beat out already, no win	○ ⑥ ④
So, what's the chance that a combo of size k , drawn from the $n - 1$ other items, is not one of the $h - 1$ other hits?	There are ten possible combinations of 2, from the 5 remaining items.	<div>①② ①③ ①⑤</div> <div>①⑥ ②③ ②⑤</div> <div>②⑥ ③⑤ ③⑥</div> <div>⑤⑥</div>
$\frac{\binom{h-1}{x} \binom{n-1-(h-1)}{k-x}}{\binom{n-1}{k}} \text{ for } x=0$	Looks like 6 out of the 10 would produce reach when #4 is added.	<div>①② ①③ ①⑤</div> <div>②③ ②⑤</div> <div>③⑤</div>
Hypergeometric: chance of x hits in a sample of k from a total of n , h of which are hits	That's 0.6	

We can use the hypergeometric formula to fill a table with a row for each possible combination size (from 0 to $n - 1$) that a hit could be added to, and a column for each possible number of total hits per respondent (0 to n), as shown in Exhibit 6 for our concrete example. Knowing that the Shapley Value is just the average contribution over all combination sizes, we can average down the columns and the column average is the Shapley Value, for each "hit" item, for a respondent with the number of hits that column is for.

Note that the column average is $1/h$ in every case⁶ (except for $h = 0$ where the SV is calculated as zero, but is irrelevant since there are no hits to which it would apply). If there are 3 hits, each has a Shapley Value of $1/3$; if there are 5, each has an SV of $1/5$. How can this be so simple?

⁶ Note also that zero entries definitely *are* included in the column averages.

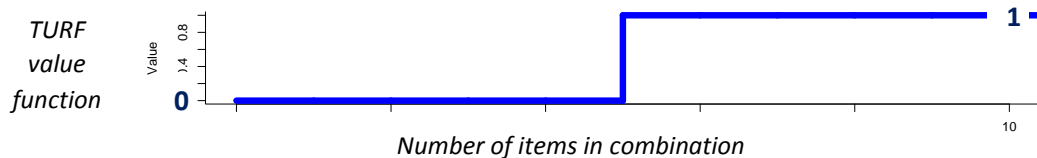
**Exhibit 6. Shapley Value contributions by combination size,
for standard TURF at the respondent level, for n=6 items.**

Each cell is $p_{\text{hypergeometric}}(x | h - 1, n - 1, k) = \frac{\binom{h-1}{x} \binom{n-1-(h-1)}{k-x}}{\binom{n-1}{k}}$ for $x = 0$

Probability a “hit” item “gets the win” (= probability of zero hits so far in combo of size k)							
Combination Size	Total Hits for a Respondent						
	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
$k = 0$	0.0	1.0	1.0	1.0	1.0	1.0	1.0
$k = 1$	0.0	1.0	0.8	0.6	0.4	0.2	0.0
$k = 2$	0.0	1.0	0.6	0.3	0.1	0.0	0.0
$k = 3$	0.0	1.0	0.4	0.1	0.0	0.0	0.0
$k = 4$	0.0	1.0	0.2	0.0	0.0	0.0	0.0
$k = 5$	0.0	1.0	0.0	0.0	0.0	0.0	0.0
Column Average	0.0	1.0	0.5	0.333	0.25	0.2	0.167

Think back to the averaging over orderings view of Shapley Values. With TURF, the value of the combination starts at zero for the null combination. As we step through an ordering, it stays zero until we get to the first hit item, whereupon the value jumps up to 1, creating an SV contribution of 1 for that first hit item. It then stays at 1, or reached, no matter what is added later. The SV contribution stays at zero except at the moment the first hit is added. See Exhibit 7. The question then is, which of our h hits will “score the win”? The first one entered will. What is the chance of a particular one being first? In all possible orderings, each has an equal shot of being the first, so it is $1/h$ for each of them.

**Exhibit 7. TURF value function as we step (left-to-right) through
an ordering of items.**



This means that we can compute Shapley Values for a sample by assigning $1/h$ to each hit at the respondent level, 0 to the non-hits, and averaging them up over the sample. Weights can easily be applied if appropriate. It is also easy to compute standard errors of the SVs, if desired. This process is exceptionally fast, scales to huge numbers of items

with no problems and is exact, not an approximation. It bypasses any version of exhaustive enumeration or evaluation of combinations.

At first glance, this may seem like a shortcut entirely tied to standard TURF. But the reasoning process behind it is far more general. Suppose we are interested in TURF at depth 3, rather than depth 1. Now an item will “score the win” if and only if there are exactly 2 hits already in the combination it is added to. So, we need only set $x = 2$ instead of 0 and create the table of hypergeometrics as before. Exhibit 8 does this. Note that the Shapley Values are, again, $1/h$, except when there are fewer than 3 hits total, making it impossible to achieve depth 3. Why? Because the item “scoring the win” is now the third hit entered in any ordering, and each hit has an equal chance of being third.

Exhibit 8. Shapley Value contributions by combination size, for depth = 3 TURF at the respondent level, for $n=6$ items.

Each cell is $p_{\text{hypergeometric}}(x | h - 1, n - 1, k) = \frac{\binom{h-1}{x} \binom{n-1-(h-1)}{k-x}}{\binom{n-1}{k}}$ for $\underline{x = 2}$

Combination Size k	Probability a “hit” item “gets the win”						
	Total Hits for a Respondent						
	$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$	$h = 6$
$k = 0$	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$k = 1$	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$k = 2$	0.0	0.0	0.0	0.1	0.1	0.1	0.1
$k = 3$	0.0	0.0	0.0	0.3	0.3	0.3	0.3
$k = 4$	0.0	0.0	0.0	0.6	0.6	0.6	0.6
$k = 5$	0.0	0.0	0.0	1.0	1.0	1.0	1.0
Column Average	0.0	0.0	0.0	0.333	0.25	0.2	0.167

We can compute such a table and the Shapley contributions for any depth, using $p_{\text{hypergeometric}}(d - 1 | h - 1, n - 1, k)$ where d is the depth of the TURF. Even without taking the shortcut directly to $1/h$, the full hypergeometric table can be easily computed in trivial time (~0.1 second for $n=200$ on a slowish laptop using R, for example).

Better still, we can extend this idea to other vaguely TURF-like value functions. Suppose we want to consider a combination successful only if it includes all a respondent’s hits, rather than just one or just d as in TURF. (This might be appropriate if the items are features and we believe a respondent will not buy a product that doesn’t include all the features she wants.) Then we can use

$p_{\text{hypergeometric}}(h - 1 | h - 1, n - 1, k)$ to populate the table, reflecting that all $h - 1$ other hits must be in a combination for the one of interest to “score the win.”

Or, suppose we will consider a combination a success if and only if it has more hit items than non-hits, a “majority” rule. Then we can use

$p_{\text{hypergeometric}}(\lfloor (k+1)/2 \rfloor \mid h-1, n-1, k)$ to populate the table, reflecting that to score the win, an item must be added to a combination that already has as many hits as half the new combination size, rounded down. The twist in this case is that adding a non-hit can take away success, so we also need to use a parallel table for non-hits, producing negative Shapley Value contributions, using

$p_{\text{hypergeometric}}(\lceil (k+1)/2 \rceil \mid h-1, n-1, k)$ as the size-specific entries.

Suppose we want a “non-excess” valuation that requires a combination to contain *only* hits, or have at most d non-hits (a depth-like notion). Here we can just reverse the original 0/1 TURF data (to 1/0), conduct a standard TURF and change the sign of the Shapley Values to negatives.

Or suppose we want a standard depth=3 TURF but with partial credit (1/3 reach for one hit, 2/3 for 2, full reach for 3 or more). We can use

$$\begin{aligned} & \frac{1}{3} p_{\text{hypergeometric}}(0 \mid h-1, n-1, k) \\ & + \frac{1}{3} p_{\text{hypergeometric}}(1 \mid h-1, n-1, k) \\ & + \frac{1}{3} p_{\text{hypergeometric}}(2 \mid h-1, n-1, k) \end{aligned}$$

giving 1/3 credit if there are no other hits in a combination being added to, 1/3 if there is one and 1/3 if there are two. All sorts of variations are possible!

In general, there is a good chance the hypergeometric shortcut will work with any value function that is “steppy” (i.e., either 0 or 1 and changing all at once as in standard TURF, or changing to only a few different values as in the depth=3 partial credit example just above, or perhaps values of -1, 0 or +1) and which treats all items equally or interchangeably.⁷ One need simply work out the combinatorial algebra and what the correct arguments for the hypergeometric probability formula are.⁸

THE BUSINESS ISSUE

The multiplicity of options for the value function highlights a major business issue as well. Standard TURF is a well-known, widely implemented value function for combinations, but it is definitely not the only option and often not the most appropriate one. Business needs, not computational convenience, should drive the selection of a value function. “Combinations” do not automatically imply “TURF”! And neither do non-TURF choices automatically imply computational problems.

⁷ Treating the items equally excludes MaxDiff data, for example, where each item has its own unique contribution rather than simply being a hit or non-hit.

⁸ In R, the *dhyper* function will handle the computations. In Excel, one must compute using the *COMBIN* function for each of the three elements of the hypergeometric probability formula. In Excel, various kinds of errors are likely in the corners and edges of the hypergeometric table, for impossible situations like the probability of two hits in a combination of size one, or the probability of two hits in a combination when the respondent only has one hit in total. *IFs* and similar conditional logic are needed to handle these; in R, *dhyper* automatically deals with those cases appropriately.

RESTRICTED-SIZE SHAPLEY VALUES

Thinking about the table of hypergeometrics suggests an interesting variant on Shapley Values, based on only some sizes of combinations. If we have 20 items, say, but are interested only in good combinations of 5 to 8 of them, why should we care how much an item contributes when added to a combination of size 18? We can easily modify the short-cut computation to average only the rows of the table we care about.

Omitting larger combinations seems an intuitively obvious idea. Whether to omit combination sizes smaller than we care about is not so clear-cut, at least to this author, but we might wish to do that as well. We might even want to compute “Shapley Values” based on a single row of the hypergeometric table, so they apply to item contributions to a single size of combination.

This sort of restriction undoes many (all?) of the mathematical axioms from which Shapley Values were derived in game theory. Out of respect for Lloyd Shapley, we shouldn’t call such things Shapley Values. But whatever the name, the idea seems like it might be useful.

In limited experimentation, the author has found that doing this changes the order and relative size of the resulting “Shapley Value-like numbers” remarkably little.⁹ That is particularly true for large n and low-depth TURF analyses, since large combinations are likely to include hits already and contributions to them are quite small. Less obviously, excluding smaller combinations also seems to have little effect, albeit more than excluding the larger ones.

Interested readers may wish to pursue this further. If so, the hypergeometric table approach makes it easy to do so.

SAMPLING COMBINATIONS

Some value functions won’t lend themselves to the hypergeometric short-cut and we are stuck working with the standard formulae. Sometimes the number of items is too large for full enumeration and evaluation. What then? We can work with random subsamples of all possible combinations. This subsection considers a few relevant details.

First, we need not feel shy about the sampling idea. There are 608 billion combinations of 18 items out of 43; it would be silly to worry about them all. We can subsample 10,000 or 100,000, be working with a far larger sample than we ever have of respondents, and still handle the computations easily.

Second, we should not simply “subsample all possible combinations.” Recall that the Shapley Value is an equally-weighted average of results for all combination sizes. Clearly, we should consider each size separately (stratify by size, if you will) to avoid the possibility of things like omitting the single null combination. There are few combinations in the small sizes, and few of the largest as well. For these, we can easily enumerate and evaluate all the possibilities without sampling.

⁹ Doing this does change the *absolute* size of the Shapley Values. Contributions for large combination sizes are often quite small; eliminating them increases the average contributions of the remaining ones. The sum of the true SVs over all items is always the maximum reach for all items together, but this sometimes-important property is destroyed by restricting the range of sizes used.

Third, we have two options as to what part of the formula we sample for. The key term in the formula is $v(\mathcal{S} \cup \{i\}) - v(\mathcal{S})$. An obvious idea would be to draw a sample of combinations of size $|\mathcal{S}|$, then add item i to each one, and evaluate both the sampled ones and the “plus i ” ones. This is statistically efficient in that we are directly sampling the *differences* that go into the Shapley Value contribution. But it is computationally messy. For each item, we need a separate sample \mathcal{S} of combinations that don’t include i . And that is true for each combination size we are sampling for. We can spend a great deal of time generating subsamples and evaluating combinations.¹⁰

A different approach is to sample a large number of combinations of size $|\mathcal{S}|$, including ones that contain item i , and a large number of size $|\mathcal{S}| + 1$, again including ones that include item i (as well as ones that do not). Then we can compute an average value for $v(\mathcal{S})$ using only the half (or so) of the combinations that don’t include i , and an average value for $v(\mathcal{S} \cup \{i\})$ using only the half (or so) of the combinations in the larger-sized sample that do include i . Instead of estimating the average difference directly, we estimate the averages and compute the difference.

The advantage of this approach is that we can generate and evaluate the samples of each size just once, not separately for every item. The computations for a given item need only subset the combinations in which it did or did not appear, take averages and subtract them. This is extremely fast. The disadvantage is that we introduce extra variance into the computation—the two samples of adjacent sizes are not matched as in the first option—so results are less accurate. The process is unbiased, however, so we can fix the variance problem by simply using larger subsamples. Even with far larger samples, say 10 times as large, overall computation speed can be far faster with the second option.

If we want to get fancy, we can combine the approaches. Consider using the first option, sampling the differences directly, with a sample size of 1,000 differences, say, for a few (equally-spaced) combination sizes and several items (chosen randomly, separately for each size), as a sort of calibration run. Compute the variances of the mean differences during the process. Then, for the same sizes and items, use samples of 10,000 per size to compute using the second option. Again, compute variances (the variance of the differences now being the sum of the variances of the two averages). Compare the variances to determine what ratio of sample sizes between options would make them about equal. Results will vary by size and item; take the largest such ratio (or at least some largish one). Apply that ratio to determine how large the samples for the second option should be to match whatever size would feel comfortable for the first option. This author would skip all this and just use ten times the sample for option 2, and probably think in terms of final sample sizes of 100,000 not being that computationally onerous. For any sizes for which there are fewer than 100,000 combinations, full enumeration would be used, of course.

A final detail concerns the mechanics of how to sample combinations. R provides functions that will generate all possible combinations, and then sample from them without replacement, all in one line of code. That seems nice, but is unworkable because

¹⁰ Clever programming can alleviate the subsample generation issue—we can sample combinations of $n-1$ items and then simply relabel items for each successive i . This does not change the number of combinations for which we must evaluate the value function, however, so is not a huge improvement.

generating all possibilities exceeds memory and CPU capabilities in the first place (in large problems). The R code accompanying this paper uses a well-known single-pass sampler to sample k of n items without replacement, thereby creating one random combination of size k , but draws 100,000 (say, or whatever number is desired) combinations that way vectorized in parallel. This is very fast, but does not prevent generation of duplicate combinations. The code then has various options to identify and drop duplicates. Usually, the best option is to ignore them. They are computationally tedious to find and drop or replace. While their presence may increase variance a tiny bit, it does not affect bias. If we sample 100,000 combinations of the 608 billion mentioned earlier, for example, there is less than a 1% chance of even a single duplicate in the first place.

MAXDIFF, TURF AND SHAPLEY VALUES

“TURF” is often applied to MaxDiff data; the quotes around TURF signal that what goes by that name is not actually TURF in the usual sense. This section discusses how that is done, how Shapley Values relate and do or do not help with that form of TURF, and what their usefulness or lack thereof in that case indicate about the underlying analysis.

As in any TURF, TURF on MaxDiff seeks to find the “best” combination of items of a given size. However, MaxDiff data is not 0/1 like TURF data is, so we must modify the standard TURF definition of combination value (i.e., the unduplicated reach) in some fashion. We will consider three general options. See Howell (2016) for a related discussion, including TURF options available in Sawtooth Software’s MaxDiff Analyzer.

DISCRETIZING THE MAXDIFF DATA

First, we can turn the respondent-level MaxDiff results into the usual 0/1 TURF data by applying some form of threshold cutoff. We might say that any item with a posterior-mean utility above some arbitrary cutoff c is a hit and the rest are not. Or we could apply the threshold cutoff to the items’ scores instead of to the utilities. Or we might base a cutoff on ranks, saying the top-scoring m items for each respondent are hits, while the rest are not.

Once this is done, we have a completely standard TURF problem, and everything discussed up to this point would apply to the TURF analysis and its Shapley Values, including the hypergeometric-based fast computation trick. This is an easy option, but it raises the question of why we bothered with MaxDiff and its scaled measurement if all we will use is a discretized 0/1 version of it. It also requires an arbitrary choice of cutoff, c or m .

WEIGHTED PROBABILITY SCORING

A second option is to say that the value function will be a “weighted probability” based on the item scores in a combination. There are various ways of transforming item *utilities* (interval-scaled, negative and positive) into *scores* (ratio-scaled, positive). All begin by exponentiating the utilities. Three important options are described below,

followed by an explanation of how to turn any of them into a weighted probability for a combination.

The simplest scoring method is to sum the exponentiated utilities across items and repercentage, so that they add to 100% or 1.00. If N is the set of all items, U_i is the utility for item i and $X_i = \exp(U_i)$ is its exponentiated utility, then the score for item i , S_i , is $S_i = X_i / \sum_{j \in N} X_j$. This “MNL approach” corresponds to calculating the probability that each item would be chosen as the best, from a choice set of all the items. No particular utility centering is required.

Another scoring option, the current default in Sawtooth Software’s MaxDiff software, is to divide each exponentiated utility by itself plus $a - 1$, so $S_i = X_i / (X_i + a - 1)$. Here, a is the number of items in the original MaxDiff tasks. This represents the probability of the item being chosen as the best from a task with a items, one being this item and the others being $a - 1$ hypothetical items of “average” strength. This approach requires that the utilities be zero-centered before exponentiating (they usually already are by default).

With anchored MaxDiff, a natural scoring is an exponentiated utility divided by itself plus 1, $S_i = X_i / (X_i + 1)$. This represents the probability of the item being chosen over the anchor (which is typically some version of none, not important, or would not buy or want), assuming that the utilities are scaled to make the anchor’s utility zero before exponentiation (which is the usual default in anchored MaxDiff).

Any of these scoring approaches for a single item can be extended into a “weighted probability” value of a combination. One simply replaces the exponentiated item utility X_i in any of them by the *sum* of the exponentiated utilities for all items in the combination. The weighted probability score represents the probability that one of the items in the combination would be chosen as best, in that scoring method’s context (i.e., vs. all other items, vs. $a - 1$ average items or vs. the anchor). The weighted probability score then becomes the value function for the combination and the “TURF” search is then for the combination of a given size with the highest weighted probability.

Do Shapley Values add anything to a TURF based on weighted probability MaxDiff scores? Consider first the MNL approach version. Here, the combination scores are the straight sum of the individual item scores¹¹. When Shapley Values are computed, the SV computation process, of subtracting the before-item combination score/value from the after-item one, exactly reverses the summation process that creates the combination scores. So, the Shapley Values are algebraically identical to the original item mean scores!

Shapley Values add nothing to our knowledge or insight in the MNL scoring case. Further, there is nothing interesting about the best combinations—if combination scores are just the sums of the item scores, then the best combination will always be that of the best individual items.¹²

¹¹ The MNL scoring case is not one of the “weighted probability” options offered in Sawtooth Software’s MaxDiff Analyzer implementation of TURF, for reasons that will become apparent. We consider it here for its expository value.

¹² Much the same is true for frequency in a standard (i.e., non-MaxDiff) TURF analysis. Neither TURF nor Shapley Values add any information beyond the original item mean frequencies. That is one reason that frequency tends to be ignored when TURF is used.

What about the average and anchor scoring approaches? In each of these, the combination scores are non-linear (but strictly monotonic and very smooth) transformations of the sums of the item scores. So, the identity between original item means and Shapley Values does not hold, and the best combinations are not guaranteed to be those of the best individual items. The continuity and non-linearity mean that the hypergeometric short-cut trick won't help us, but we can always compute SVs by brute enumeration or sampling, depending on problem size.

But, consider the shape of these non-linear transformations of total item scores, as illustrated in Exhibit 9.¹³ What we see is that while the average and anchor weighted probability scores are not linear, they are not that far from it.

This implies that Shapley Values computed from them will be approximately proportional to the original item mean scores. And that implies that, as in the MNL scoring case, the Shapley Values will add little new information. Perhaps more importantly, it implies that the best-scoring combinations will rarely be anything other than the best individually-scoring items. That, of course, calls into question the entire value of any weighted probability approach to TURF on MaxDiff.

RESTORING THE THRESHOLD IN TURF

A third way to implement TURF for MaxDiff data is to calculate the weighted probability scores as above, but not use them as the final value function. Instead, we return to the reach/non-reach “threshold” idea of TURF by establishing some cutoff d , and saying that a combination achieves reach (i.e., has a value function of 1, as opposed to 0) if the weighted probability exceeds the cutoff.

Exhibit 10 illustrates the resulting value function, based in this case on the anchored scoring curve (it will work similarly with any of the three scoring approaches discussed, and others as well). Note that it restores the curve shape illustrated back in Exhibit 7. As with the weighted probability approach without a threshold, computing Shapley Values in this situation requires enumeration or sampling, with no help from the fast hypergeometric short-cut.

¹³ The vertical scales in Exhibit 9 are different for each curve, to allow us to approximately superimpose the curves and compare their shapes, independently of their general slopes.

**Exhibit 9: MaxDiff Weighted Prob. Scoring:
Sum of Item Scores vs. Combo Scores**

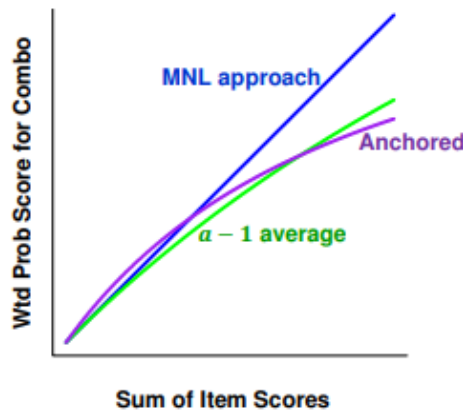
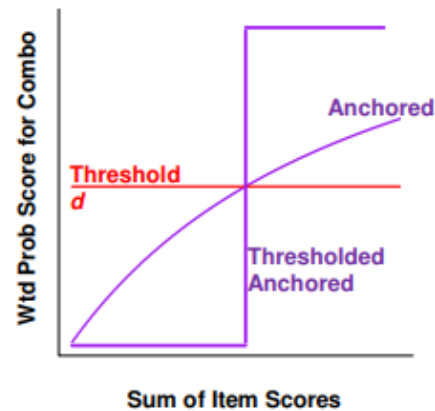


Exhibit 10: Threshold on Wtd. Prob. Score



Like the straight weighted probability approach, this has the advantage of preserving and respecting the original continuous MaxDiff data. But it is also more like classic TURF in that “reach” is either achieved or not. It does require an arbitrary choice of cutoff d , but that choice is closely analogous to choosing the depth in a standard TURF, and not unnatural in a TURF context.

In this author’s opinion, this is the only approach for which the term “TURF on MaxDiff” is particularly appropriate. We might call the discretization approach “TURF on ‘MaxDiff’”—real TURF, but on severely degraded MaxDiff data. We might call the straight weighted-probability-as-value approach “‘TURF’ on MaxDiff”—it respects the MaxDiff data but uses a value function with little resemblance to standard TURF.

CONCLUSIONS: TURF AND SVs ON MAXDIFF DATA

As just stated, only some form of thresholding on some version of a weighted probability score seems worthy of the term TURF on MaxDiff. But, does that mean that is the right approach to use? Not necessarily.

The key issue in choosing a value function is not whether it is worthy of some label, but whether it makes sense in the actual business case. Does the thresholding feature of TURF make sense? Sometimes yes, but quite often no.

There is natural appeal to the idea that we should devote resources to raising some consumers above some minimum standard or barrier before we further impress, satisfy or delight those who are already there. From that standpoint, thresholding makes sense. On the other hand, saying that improvements are worthless unless they drag us across some arbitrary line makes little sense. (Note that that is one of the prime criticisms of Net Promoter Scores.) From the latter viewpoint, weighted probability scoring without a threshold is a perfectly reasonable idea. In the original TURF application of media exposures, exposed-or-not is a clear-cut distinction. In many modern marketing applications of TURF there is no bright line, so a “reach” threshold may be entirely artificial. This is often true for TURF on standard 0/1 data as well and is only partly cured by looking at multiple depths of analysis.

If we do think a weighted probability value (i.e., without threshold) is appropriate, thinking about the behavior of Shapley Values in that case does suggest that TURF analysis will add little insight, as the value of any combination generally tracks the total value of its component items. And the Shapley Values themselves will add little beyond the original mean item scores. Simply working with original item scores may be the best course.

SUMMARY

Shapley Values can be easily understood as average item contributions over orderings, a useful paradigm in the context of things like key driver regressions that have a natural ordering interpretation. They are also average item contributions to combinations, with all combination sizes weighted equally, a useful viewpoint in the context of TURF and many common marketing research applications.

Their usefulness is in reducing a sea of combination tallies to a manageable summary per item, providing a more compact, insightful and memorable overview of what is going on with the combinations. Even when a list of best combinations is what is ultimately needed, the Shapley Values can provide a helpful road map to the data. Shapley Values are also more stable than TURF results, facilitating tracking and subgroup comparisons.

Computation of Shapley Values need not be daunting. Small problems can be brute-forced easily. Large ones can always be attacked with sampling of combinations. In many TURF-like cases, an exact, ultra-fast computation can be done.

A central idea in applying Shapley Values is that of the value function—how do we quantify the “goodness” of a combination? Unduplicated reach as in TURF is a very common answer, but by no means the only one and very often not the best one. The Shapley Value idea applies to *any* value function (so does the TURF notion of searching for the best possible combination). It is crucial to consider the underlying *business issue* when deciding what value function makes sense.

Thinking about how Shapley Values behave in the context of TURF on MaxDiff data using weighted probability approaches suggests that they add little, and in fact that such TURF on MaxDiff approaches are unlikely to reveal very much. Using a thresholded form of value function would help that problem, but to reiterate, it is the *business issue*, not the analytical details, that should decide how we value a combination.

About half the conference attendees indicated they had used Shapley Value analysis at least once. That is more than this author had expected, but he believes they could be even more widely useful and hopes this paper will facilitate and encourage that use.



David Lyon

APPENDIX: SOFTWARE FOR SHAPLEY VALUES, AND TURF

The R package *relaimpo* will do Shapley regression, as for key driver regressions. Use type=“lmg” for average r-squared contributions. Although efficiently and strategically implemented, it does evaluate all possible combinations (not orderings), so run times become unreasonable on a PC after 25 or fewer items, and roughly double with each additional item.

R package *turfR* by Jack Horne will perform TURF analyses, for not-too-large problems, but does not do Shapley Values. The *turfR* code is faster than that provided with this paper, but at the cost of very high memory use that limits the problem size. A different, earlier, R package named simply *turf* is not particularly useful.

The author’s R code was distributed at the conference with the slide handouts and is available from the Sawtooth Software website at <http://sawtoothsoftware.com/download/lyon2018.zip>, or by email request to the author at dlyon@aurora2000.com. (The version with the first handout distribution was incomplete; later versions were OK, but downloading it from the URL will guarantee having the most up-to-date version.) It includes Shapley Value computation routines using full enumeration and evaluation, and others using sampling of combinations, that work with any user-supplied value function (referred to as “a scorer” in the code comments). It also includes code to implement the hypergeometric fast computation approach for TURF of any depth, readily modifiable for many other value functions. It also includes a scorer for TURF of arbitrary depth, one for coverage, and some MaxDiff-relevant routines. All these routines generally handle weighted data.

REFERENCES

- Conklin, Michael and Stan Lipovetsky (2000), “A New Approach to Choosing Flavors,” Advanced Research Techniques Forum presentation, Monterey.
- Conklin, Michael and Stan Lipovetsky (2005), “Marketing Decision Analysis by TURF and Shapley Value,” International Journal of Information Technology & Decision Making, volume 04, pp. 5–19.

- Conklin, Michael and Stan Lipovetsky (2013), “The Shapley Value in Marketing Research: 15 Years and Counting,” Proceedings of the 2013 Sawtooth Software Conference.
- Conklin, Michael and Faina Shmulyian (2012), “Portfolio Management: Combining DCM and Shapley Value Line Optimization,” Advanced Research Techniques Forum presentation, Seattle.
- Cortese, Ben (2018), “Bayesian Network Key Driver Analysis,” Proceedings of the 2018 Sawtooth Software Conference (this volume).
- Howell, John (2016), “A Simple Introduction to TURF Analysis,” Sawtooth Software technical paper.
- Johnson, Jeff and James Lebreton (2004), “History and use of relative importance indices in organizational research,” *Organizational Research Methods*, vol. 7, pp. 238–257.
- Kruskal, William (1987), “Relative Importance by Averaging over Orderings,” *American Statistician*, February 1987, pp. 6–10.

FDA SEEKS PATIENT PREFERENCE INFORMATION TO ENHANCE THEIR BENEFIT-RISK ASSESSMENTS: CASE STUDIES

LESLIE WILSON

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

JORDAN LOUVIERE

UNIVERSITY OF SOUTH AUSTRALIA

The inclusion of the patient voice in shared health care decision-making has progressed from the physician's office to the regulatory approval of medical devices. The FDA Patient Preference Initiative is expanding the need for patient preference research and the frameworks that can support the conduct of discrete choice experiments in the health care market. It is important to understand the growth of patient preference research within the FDA and to understand how their design needs are developing through exploring a few examples of the use of patient preference in FDA approval decisions.

The goal of this paper is to describe the landscape for patient preference studies in FDA regulatory decision making and to present case studies of different discrete choice methods of patient preference that are used and planned for use in FDA regulatory decisions for medications and devices.

FDA LANDSCAPE: THE PATIENT PREFERENCE INITIATIVE

Recent amendments to the Prescription Drug User Fee Act (PDUFA) required the FDA to include patient preference in its structured benefit risk framework. In 2013, the FDA launched the Patient Preference Initiative to incorporate patients' views as scientific, empirical evidence when appropriate, to their decisions (US Food and Drug Administration). To implement this development, the FDA Center for Devices and Radiological Health (CDRH) and Center for Biologics Evaluation and Research (CBER) first collaborated with the Medical Device Innovation Consortium (MDIC) to develop a framework report "A Framework for Incorporating Information on Patient Preferences Regarding Benefit and Risk into Regulatory Assessments of New Medical Technology" published in May 2015 (Medical Device Innovation Consortium, 2015). This framework included a catalogue of methods for assessing patient preference, an analysis of gaps in current assessment methods and an agenda for further research and was the basis for the CDRH/CBER publication of a draft guidance effective from October 2016 (Center for Devices and Radiological Health, 2016). This FDA guidance defined patient preference information as the "qualitative or quantitative assessments of the relative desirability or acceptability to patients of specified alternatives or choices among outcomes or other attributes that differ among alternative health interventions" and described different approaches currently available to quantify and collect patient preference information (U.S. Department of Health and Human Services #1, 2016). These approaches included quantitative discrete choice measures such as choice-based conjoint analysis and best-worst approaches. CDRH also began collaborations with preference researchers to initiate case examples of patient preference for devices considered "preference-sensitive."

Preference sensitive conditions are those where clinical evidence does not support a single option and the appropriate options depend on the values or preferences of the beneficiary (U.S. Department of Health and Human Services #1, 2016).

The FDA regulatory division for drugs, Center for Drug Evaluation and Research (CDER), initiated a more qualitative approach to including preference into their regulatory decisions. They initiated their Patient-focused Drug Development (PFDD) approach with a goal to better incorporate the patient's voice in drug development and evaluation. Their efforts included FDA-led disease-specific PFDD meetings to obtain the patient perspective. To date they have conducted and posted reports on 22 of these public meetings (US Department of Health and Human Services #2).

Finally, the CDRH, CBER, and CDER supported a workshop with the five Centers of Excellence in Regulatory Science and Innovation (CERSI) centers titled "Advancing Use of Patient Preference Information as Scientific Evidence in Medical Product Evaluation" (US Department of Health and Human Services #3). The CERSI centers are collaborations between Academic Institutions and the FDA to advance regulatory science through innovative research, education and scientific exchanges and are acting to support the FDA's patient preference initiative (University of California, San Francisco).

The pharmaceutical industry also is exploring how they can incorporate patient preference into their drug development process with the initiation of a public private partnership called PREFER. This is a five-year research project to assess when and how patient preference on benefits and risks should be incorporated into decisions on medicinal products. This initiative has 3 parts: Part A: Literature reviews and interviews to gain insights from all stakeholders on needs and methods, Part B: Testing preference-elicitation methods in clinical case studies, and Part C: Developing recommendations and guidelines for design, conduct, analysis, and reporting of patient preference studies for industry, regulatory authorities and HTA bodies around patient preference (PREFER, 2017).

The FDA is encouraging the use of patient preference information throughout the health care and product life cycle beginning with device developers and patient groups and extending to manufacturers and FDA regulators. All of these developments can result in an explosion of discrete choice patient preference experiments in the health care marketplace. The following case studies can provide examples of how patient preference experiments for healthcare decision making around risk and benefit rather than price are unique in sampling, design, and attribute selection.

CASE STUDIES

We present case studies of three health interventions where patient preference was used in FDA decision making, the methods used, their results and the impact on FDA decision making: 1) Treatments for Multiple Sclerosis, 2) Duchenne's Muscular Dystrophy (DMD), and 3) EnteroMedic's Maestro obesity device. We will also describe the development of a choice-based conjoint measure for use in future CDRH decision making for Upper-Limb Loss Prosthetic Devices in collaboration with the FDA and describe its focus on validity testing of discrete choice measures.

Multiple Sclerosis and Strength of Patients to Recall a Drug

One of the first examples of the ability of patients' views to affect regulatory decisions was for disease modifying therapies (DMTs) for patients with relapsing-remitting multiple sclerosis (RRMS). MS is an autoimmune disease affecting about 400,000 people in U.S. After a long period with little advance in MS treatments from the use of alpha and beta interferons (Durelli L. et al., 1995), a new drug, Natalizumab (TysabriTM), was approved in November 2004. This drug had a dramatically better clinical outcome for RRMS, including a 42% reduced risk of disability progression, improved cognitive performance and 66% fewer relapses (Klawiter E.C. et al., 2009). Some patients' response to this treatment allowed them to stop using a wheelchair for the first time. After just 4 months on the market, however, there were rare documented cases of progressive multifocal leukoencephalopathy (PML) a CNS disorder which results in severe disability or death (Wenning W. et al., 2009) and in consultation with the FDA, the manufacturer voluntarily withdrew the drug from the market in February 2005. However, this resulted in a change of treatment for many patients who were currently appreciating the full and sometimes dramatic benefits of the new treatment. Through their patient advocacy organization (the National MS Society, NMS), these patients voiced their concerns about the sudden drug withdrawal, suggesting that for some the benefits were worth the risks and that each patient should be able to weigh these risks and benefits for themselves. NMSS commissioned a national survey to probe the level of risk 810 people with MS were willing to take in their use of Tysabri if it went back on the market. Opinions were evenly distributed from very positive to very negative, with half offering no definite opinion about the drug's return to the market and ranged from immediately to more than 1 year in how long they would wait to use it (MS Society, March 8, 2006). This public pressure, however, led to the return of Tysabri to the market in March under a "Black-box" warning of the risks of PML and a requirement for patient registration into a program to inform patients of the risks and to ensure safe use of the drug.

Only one other prescription drug has ever returned to market after being pulled because of dangerous side effects, making RRMS treatments one of the earliest preference sensitive conditions with large benefits, and very small but devastating risks where regulators listened to the patients' voice. In the meantime, there were many patient preference studies conducted to begin to document patients' risk-benefit trade-offs (Shingler S.L. et al., 2013; Johnson F.R. et al., 2009; Paulos C. et al., 2016; Utz K.S. et al., 2014; Rosato R. et al., 2015; Wicks P. et al., 2015). Wilson et al. published two papers describing RRMS patient preferences across the full range of DMTs. They found that patients were willing to accept 0.08% severe risk of death or severe disability for a year delayed relapse, and 0.22% for a 4 versus a 2 year prevention in progression (Wilson L. et al., 2014 and Wilson L. et al., 2015). Interestingly, patients indicated that they were willing to accept more risk than actually was demonstrated by the current DMT PML risk (1/1,000), depending on the benefit gained. They also found that how patients feel (i.e., symptom improvement) was the most preferred among all other DMT benefit attributes studied, despite this not being a proven benefit of clinical trials at the time.

All of these patient preference studies were helpful in demonstrating that patients' views were variable and often were less risk averse than the FDA and physicians expected. Currently more is known about PML and patients can be tested to better

identify a more personalized risk profile. In addition, more DMTs with different risk profiles are now available to choose from. This MS case study, though, is the first and an important example of how the FDA, disease societies and patients can learn from each other once the patient voice is examined, especially when using quantitative methods of discrete choice surveys.

Duchenne's Muscular Dystrophy: The Strength of Patient Advocacy

Duchenne's Muscular Dystrophy (DMD) is a genetic disorder characterized by progressive muscle weakness and degeneration. It is caused by a mutation in the gene encoding the dystrophin protein which is critical to muscle integrity. The onset is primarily in males beginning around 3-5 years of age and most die in their 20s. Current treatment is corticosteroids, with no specific FDA approved treatment for DMD. The FDA (CDER), through the Patient Preference Initiative, is conducting public Patient Focused Drug Development meetings for 20 conditions to obtain patients' views on living with their illness, the symptoms, the treatments, their values for living with the disease and their willingness to join a clinical trial. DMD was not one of the diseases chosen nor were other relatively rare diseases. Therefore, the Parent Project Muscular Dystrophy (PPMD) advocacy organization partnered with patient preference researchers to conduct a patient preference study to complement these meetings. Peay et al. (2014) conducted a Best-Worst Scaling caregiver preference study to explore their preferences for emerging treatments for DMD and to highlight principles of patient-centered outcomes research with an advocacy organization's leadership (Peay et al., 2014). Among the 119 DMD caregivers, treatment effect on muscle function, risk of heart arrhythmia, and risk of bleeding were the most important attributes and having additional post approval data was the least important variable. This demonstrated their views of the importance of promoting patient-centered drug development with shorter development times, and willingness to accept unknown risks for the ability to try an unproven treatment.

Hollin I.L. et al. (2015) conducted a follow-up study comparing two stated-preference methods, best-worst scaling (BWS) and conjoint analysis (CBC) applied across DMD's potential treatments (Hollin I.L. et al. 2015). The BWS attributes were 1) speed of progression of weakness, 2) gain in lifespan, 3) amount of post-approval drug information available, 4) loss of appetite, 5) increased risk of bleeding, and 6) increased risk of heart arrhythmia. They found that those affected by life threatening and debilitating illness are willing to accept risks and uncertainty about those risks (Hollin I.L. et al., 2015). They also demonstrated that the BWS and CBC approaches gave similar preference results ($p < 0.01$). The CBC results demonstrated that patients were willing to exchange high probabilities of side effects and additional blood draws to maintain cough strength for 10 years (Hollin I.L. et al., 2015). In addition to these quantitative studies demonstrating patient preferences, the patient advocacy group also developed the first proposed draft guidance document for industry for submission to the US Food and Drug Administration. The FDA embraced this work and collaboratively the FDA and CBER published this DMD Final Guidance for Developing Drugs for Treatment on February 2018 (US Department of Health and Human Services #4). The goal of this Guidance was to assist drug companies in the clinical development of drugs

for the treatment of DMD and related diseases. This is the first time, however, that a proposed draft guidance was independently prepared by an advocacy group and shows the strength of patient advocacy to use discrete choice measurement to affect drug development in a new way.

Implanted Devices to Promote Weight Loss: Use of Patient Preference in Device Approval Decisions

A third case example, for a vagus nerve gastric stimulator device surgically implanted for weight loss in “obese” subjects, demonstrates the value of patient preference information as a primary factor in the approval process for devices. Drs. Ho et al. (2015) conducted a CBC in 500 obese patients from an online panel representative of the demographics of the U.S. obese population, stratified by body mass index (BMI) (Ho M.P. et al., 2015). They selected and carefully defined eight attributes descriptive of the risks and benefits of all types of surgically implanted bands and pilot tested the attributes with face-to-face interviews. Risk attributes included mortality, adverse events, and need for hospitalization, while benefits included relative weight loss amount and duration, and improvement in comorbidities associated with obesity. Other key attributes were type of surgical procedure and diet restrictions required with the device use. They found that patients were willing to trade off a 0.01% mortality risk for a 10% total body weight loss lasting for 5 years (Ho M.P. et al., 2015). This CBC information was used by regulators as primary evidence to make the approval decision for the EnteroMedics’s Maestro Rechargeable System implantable device. This device is unique in electrically stimulating the vagus nerve to indicate to the brain that the stomach is full, compared with the other two weight loss devices the FDA has approved, Lap-Band Gastric Banding System, and Realize Gastric Band which both physically restrict the ability of the stomach to contain food. The Maestro device trial demonstrated safety, but did not meet its primary endpoint to reach a 10% difference in weight loss at 12 months compared to the sham control group, but was approved anyway. Approval of this device was therefore based in large part on the patient preference results which demonstrated that a large portion of obese patients would accept the risks associated with a surgically implanted device if they lost a sufficient amount of weight. This was the first quantitative patient-preference study designed and used to support a regulatory approval decision by FDA Center for Devices and Radiological Health.

Based on this work, researchers also developed an online study tool to define minimum clinical effectiveness that can be used to inform future benefit-risk assessments for other pre-market approvals of weight-loss devices. Determination of the “minimum clinical effectiveness” value is especially important to regulatory reviewers as it is used when designing clinical studies to both size the studies and to decide whether the benefits of the treatment outweigh the risks for market approval. They used their CBC results to build a MaxR-MinB calculator that could be used across weight-loss devices to help CDRH reviewers determine MaxR and MinB of an average patient as well as an early adopter of a device that provides a given weight loss and poses a mortality risk. Given this type of information, the FDA stated that they might consider approving a device only for risk-tolerant patients, and indicate these limits in the device label. This study provided a “proof of principle” to support FDA Guidance documents about the use of patient

preference evidence for device decisions. This example is likely to lead to more demand for patient preference information to support device development and approval.

Prosthetic Devices for Limb Loss: Preparing for Approval of the First Implantable Components for Prosthetic Devices

The FDA Guidance on patient preferences for use in approval decisions and the success of their first CBC study for the gastric weight-loss device encouraged the CDRH to continue case studies to support continued use of qualitative preference measurement approaches. But CDRH recognized that they also needed information on the validity of these methods. They provided grant support to Wilson L. through the UCSF/Stanford CERSI center to examine the validity of different preference measurement approaches in a preference sensitive condition that was experiencing fast innovation, limb prosthetics. Two main innovations were being developed for the first implantable prosthetic components. These devices were preference sensitive because up to 70% of patients with upper limb loss reported not using their prosthetics despite their initial adoption (Ziegler-Graham K. et al., 2008; Raichle K.A. et al., 2008). It is essential to know how patients weigh risks and benefits of new prosthetic innovations for regulatory decisions.

The two innovations that received funding for quantitative preference studies were osseointegration and myoelectric control. Osseointegration is a new prosthetic attachment technology which surgically implants a titanium post into the bone which then anchors by growth of bone and tissue around it. The device eliminates the need for a heavy shoulder harness and socket and its problems with fit and skin abrasion, and allows a better range of motion and enhanced feeling of device integration. However, osseointegration implantation also requires two surgical procedures and a continual risk of infection around the post. Myoelectric control devices are implanted electrodes which detect minute muscle, nerve and EMG activity to control prosthetic limb movements. They offer more natural and accurate motions, but also add weight, may require surgical implantation of sensors, and require substantial training for successful use. Because there is no established regulatory paradigm for either of these devices, they are an ideal case study for preference assessment and validation.

We worked with the FDA, prosthetists, prosthetics device developers and upper limb loss patients to select the 10 most important attributes for these devices. We used a modified meta-ethnography approach to select the attributes and define them conceptually. This technique involves a process of sorting relevant literature and patient findings/statements into patterns of evidence and evaluating their importance at deeper and deeper conceptual levels. We pilot tested our attribute selections in pairs of specialists of different stakeholders which allowed us to slim the list of attributes to 9 and to clearly define them and the 3-4 risk and benefit levels of each. Our risk attributes were need for surgery, risk of infection, probability of experiencing daily pain, and risk of complete loss of prosthetic use. Beneficial attributes included improved grip, improved range of motion, ability to feel sensations, and feelings of integration with the prosthetic. We also developed both a CBC measure without video demonstrations and one with video to demonstrate the motion ability of some of the attributes. We are comparing these two CBC approaches in a 25-patient convenience sample at UCSF as a pilot test.

Preliminary analysis of our first 10 subjects with upper limb loss demonstrates that anecdotally patients prefer the video descriptions to the plain text labels. The utilities in the CBC with video show a stronger utility and disutility than the same CBC attributes without video. In addition preliminary analysis of each attribute's preference score (calculated as % chosen/% shown) demonstrates that pain and infection risk are the least preferred attributes, while hand grip patterns, strength and the ability to independently cook dinner were the most preferred attributes. Subjects showed that they were willing to trade risks for benefits. Subjects were not able, however, to identify any attributes that could be eliminated to reduce the cognitive burden. More accurate analysis will be performed after we recruit more subjects. In addition, through additional funding from the Burroughs Wellcome Fund, we will be extending this research to include testing of two types of validity for use of CBC for regulatory purposes: concurrent validity and convergent validity. Concurrent validity will be examined by comparing three different measurement approaches; CBC with video use, CBC without video use and Standard Gamble utility measurement. We will test convergent validity by making utility comparisons between those with single upper limb loss and bilateral upper limb loss as well as those with loss on the dominant side vs. the non-dominant limb; expecting that those with bilateral limb loss and loss on the dominant side will be more risk averse than those with unilateral loss. Finally, an important question for all discrete choice researchers and users, and especially for those using CBC to make regulatory decisions in health care, is whether or not subjects' stated preferences actually reflect their revealed preferences. We will address this question in our research plan, by comparing subjects' CBC scores before and after they undergo an osseointegration procedure for lower limb loss. These results will provide further support to the FDA for use of CBC in regulatory decisions.

CONCLUSION

The FDA is seeking patient preference studies that can serve as case examples to further advance their goal of including the patient voice in regulatory decisions for both drugs and devices (Marshall D. et al., 2010; Hall J. et al., 2004; Louviere J. et al., 2000; Johnson F.R. et al., 2016). We describe previous case examples and how CBC and other patient preference measurement techniques are being evaluated as a useful tool for risk benefit decisions of regulatory FDA bodies. CBC studies performed to assist in regulatory decisions differ from other uses of CBC in several ways. First, they cannot include factors of price/cost because this cannot be part of the FDA approval decisions for medicines or devices. Additionally, the CBC attributes generally must be applicable across products rather than specific to one product. Positively, patient subjects are generally very invested in the CBC process, what is being asked and the importance of giving their opinions, which means that engaging them in the process may be easier than for marketing studies. Validity for discrete choice experiments has been primarily tested for non-health care examples (Louviere J. et al., 1992; Menictas C., Wang P.Z., Louviere J., 2012). Finally, because of the need to ensure patient safety establishing the validity of these methods is even more important. Although work still is needed to identify the most valid methods of measuring patient preference for regulatory decisions, there is general agreement that including the patient voice is essential to making these difficult decisions of what treatments are safe and effective.



Leslie Wilson



Jordan Louviere

REFERENCES

1. US Food and Drug Administration, USDHHS,
<https://www.fda.gov/aboutfda/centersoffices/officeofmedicalproductsandtobacco/cdrh/cdrhpatientengagement/ucm462830.htm>. Accessed August 3, 2017.
2. Medical Device Innovation Consortium (MDIC) Patient Centered Benefit-Risk Project Report. 2015 http://mdic.org/wp-content/uploads/2015/05/MDIC_PCBR_Framework_Web1.pdf. Accessed August 3, 2017.
3. Center for Devices and Radiological Health and Center for Biologics Evaluation and Research. Guidance for Industry and Food and Drug Administration Staff: Factors to Consider When Making Benefit-Risk Determinations in Medical Device Premarket Approval and De Novo Classifications. Issued on October 23, 2016. Available from: <https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm517504.pdf>. [Accessed July 28, 2017].
4. U.S. Department of Health and Human Services #1., Food and Drug Administration, Center for Devices and Radiological Health and Center for Biologics Evaluation and Research. Patient preference information—submission, review in PMAs, HDE applications, and De Novo requests, and inclusion in device labeling: draft guidance for industry, Food and Drug Administration staff, and other stakeholders. October 23, 2016. Available from: <https://www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446680.pdf>. [Accessed July 28, 2017].
5. US Department of Health and Human Services #2. US Food and Drug Administration. Patient-Focused Drug Development: Disease Area meetings held in Fiscal Years 2013–2017. <https://www.fda.gov/ForIndustry/UserFees/PrescriptionDrugUserFee/ucm347317.htm> Accessed August 3, 2018.
6. US Department of Health and Human Services #3. US Food and Drug Administration. Advancing the use of patient preference information as scientific evidence in medical product evaluation.

<https://www.fda.gov/ScienceResearch/SpecialTopics/RegulatoryScience/ucm574320.htm>. Accessed August 3, 2018.

7. University of California San Francisco. Schools of Pharmacy and Medicine, Department of Bioengineering and Therapeutic Sciences. UCSF-Stanford Center of Excellence in Regulatory Science and Innovation (CERSI). <https://pharm.ucsf.edu/cersi>. Accessed August 3, 2018.
8. PREFER. Patient Preferences. Patient preferences in healthcare decision-making 2017. <https://www.imi-prefer.eu/news/news-item/?tarContentId=642846>. Accessed August 3, 2018.
9. Durelli L., Bongioanni M.R., Cavallo R., Ferrero B., Ferri R., Verdun E., Bradac G.B., Riva A., Geuna M., Bergamini L., Mult Scler., et al. 1995;1 Suppl 1:S32-7. Interferon alpha treatment of relapsing-remitting multiple sclerosis: long-term study of the correlations between clinical and magnetic resonance imaging results and effects on the immune function.
10. Klawiter E.C., Cross A.H., et al. Neurology. The Present Efficacy of Multiple Sclerosis Therapeutics. 2009. 73(12): 984–990.
11. Wenning W., Haghighi A., et al. Treatment of progressive multifocal leukoencephalopathy associated with Natalizumab. NEJM. 2009. 361:1075–1080.
12. MS Society. March 8, 2006. People with MS surveyed for Views. https://secure.nationalmssociety.org/site/SPageServer/?NONCE_TOKEN=F134B1DB91A17F9773C8103FB557C04E&pagename=HOM_RES_tysabri_surveyresults. Accessed August 3, 2018.
13. Shingler S.L., Swinburn P., Ali S., Perard R., Lloyd A.J. A discrete choice experiment to determine patient preferences for injection devices in multiple sclerosis. J Med Econ. 2013; 16:1036–1042.
14. Johnson F.R., Van Houtven G., Ozdemir S., et al. Multiple sclerosis patients' benefit-risk preferences: serious adverse event risks versus treatment efficacy. J. Neurol. 2009; 256: 554–562.
15. Poulos C., Kinter E., Yang J.C., Bridges J.F., Posner J., Reder A.T. Patient preferences for injectable treatments for multiple sclerosis in the United States: a discrete-choice experiment. Patient. 2016; 9:171–180.
16. Utz K.S., Hoog J., Wentrup A., et al. Patient preferences for disease modifying drugs in multiple sclerosis therapy: a choice-based conjoint analysis. Ther Adv Neurol Disord. 2014; 7:263–275.
17. Rosato R., Testa S., Oggero A., Molinengo G., Bertolotto A. Quality of life and patient preferences: identification of subgroups of multiple sclerosis patients. Qual Life Res. 2015; 24:2173–2182.
18. Wicks P., Brandes D., Park J., et al. Preferred features of oral treatments and predictors of non-adherence: two web-based choice experiments in multiple sclerosis patients. Interact J Med Res. 2015; 4: e6.

19. Wilson L., Loucks A., Bui C., et al. Patient centered decision making: use of conjoint analysis to determine risk-benefit trade-offs for preference sensitive treatment choices. *J Neurol Sci.* 2014; 344: 80–87. 10.
20. Wilson L.S., Loucks A., Gipson G., et al. Patient preferences for attributes of multiple sclerosis disease-modifying therapies: development and results of a ratings-based conjoint analysis. *Int J MS Care.* 2015; 17: 74–82.
21. Peay, H.L., Hollin I., Fischer R., Bridges J.F.P. Clinical Therapeutics/Volume 36, Number 5, 2014 A Community-Engaged Approach to Quantifying Caregiver Preferences for the Benefits and Risks of Emerging Therapies for Duchenne Muscular Dystrophy.
22. Hollin I.L., Peay H.L., Bridges J.F. Patient. 2015 February; 8(1):19–27. Caregiver preferences for emerging duchenne muscular dystrophy treatments: a comparison of best-worst scaling and conjoint analysis.
23. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Devices and Radiological Health and Center for Biologics Evaluation and Research. Duchenne muscular dystrophy and related dystrophinopathies: developing drugs for treatment guidance for industry draft guidance. June 2015. Available from: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM450229.pdf>. [Accessed July 28, 2017].
24. Ho M.P., Gonzalez J.M., Lerner H.P., Neuland C.Y., Whang J.M., McMurry-Heath M., Hauber A.B., Irony T. Incorporating patient-preference evidence into regulatory decision making. 2015. *Surg Endosc* 29(10); 2984.
25. Ziegler-Graham K., MacKenzie E.J., Ephraim P.L., Travison T.G., Brookmeyer R. Estimating the prevalence of limb loss in the United States: 2005–2050. *Arch Phys Med & Rehab.* 2008;89(3):422–429.
http://biomed.brown.edu/Courses/BI108/BI108_2003_Groups/Hand_Prosthetics/stats.html; <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC31936371>
26. Raichle K.A. et al. Prosthesis use in persons with lower- and upper-limb amputation. *J Rehabil Res Dev.* 2008; 45(7):961–972.
27. Marshall D., Bridges J.F., Hauber B., et al. Conjoint analysis applications in health—how are studies being designed and reported? An update on current practice in the published literature between 2005 and 2008. *Patient* 2010;3:249–56.
28. Hall J., Viney R., Haas M., Louviere J. Using stated preference discrete choice modelling to evaluate healthcare programs. *J Bus Res* 2004; 57:1026–32.
29. Louviere J., Hensher D., Swait J. Stated Choice Methods: Analysis and Applications. Cambridge University Press, Cambridge, UK, and New York, NY, 2000.
30. Johnson F.R., Ma M.Z. Patient Preferences in Regulatory Benefit-Risk Assessments: A US Perspective Value in Health. 2016; 19:741–45).

31. Louviere J.J., Timmermans H.F.P. Testing the External Validity of Hierarchical Conjoint Analysis Models of Recreational Destination Choice. *Journal Leisure Sciences*. 14:(3) 179-194.
32. Menictas C, Wang P.Z., Louviere J.J. 2012. Assessing the Validity of Brand Equity Constructs, *Australian Marketing Journal*, 20(1) 3-8.

A DIRECT COMPARISON OF DISCRETE CHOICE AND ALLOCATION CONJOINT METHODOLOGIES IN THE HEALTHCARE DOMAIN

JAMES PITCHER
TATIANA KOUDINOVA
DANIEL ROSEN
GfK

1. ABSTRACT

Patient Based Discrete Choice (PBC) and Allocation Based Conjoint (ABC) are both commonly used to estimate new product preference shares in the healthcare space. For the first time, this research directly compares the accuracy of the two methods, their characteristic similarities and differences, as well as their ease of implementation and respondent-friendliness. Our research revealed significant differences between the two models both in terms of modelled preference share estimates and directly reported preference share.

2. BACKGROUND

GfK commonly uses two distinct methodologies to estimate new product preference shares in the healthcare and pharma space, the first of which asks physicians to report prescribing preferences for specific real world patients, and the second of which asks physicians to report their prescribing decisions at a practice level rather than on a per-patient basis. This distinction reflects a trade-off market researchers often make when designing a research study, whether to have the research environment closely resemble the real world decision environment, or whether a carefully designed, albeit “artificial,” research environment elicits more accurate information from respondents.

It is generally beneficial to have one’s statistical toolbox stocked with multiple methodologies, but in this case there is a lack of clear guidance as to which method is best (or in which situations one method is preferred over another). There is currently no fact-based evidence to guide method selection, nor is there any third party literature directly related to this problem. The result is that method selection is not determined by empirical research, but simply by the historical experience and comfort level of the research teams.

In this study we attempt to put method selection on an evidenced based track by comparing two specific research methodologies in order to understand their relative predictive accuracy, differential characteristics and comparative user-friendliness. The two methods studied are:

Patient Based Discrete Choice (PBC):

In this method physicians are asked to consider real patient cases. They then are presented with treatment options and answer a set of questions on their product preferences and estimated behavior in regards to these patients, taking into account all patient characteristics that could play a role in their decision making.

Allocation Based Conjoint (ABC):

In this method physicians are also presented with a handful of treatment options, except instead of being asked to choose their most preferred treatment for a particular patient, they are asked to imagine the treatment choices they would make for their next group of patients. (We typically define a group as 10–100 patients.) Physicians then report how many of these patients they would choose to treat with the first treatment option, how many with the second, and so on. This is called an allocation exercise since different numbers of patients are allocated to each of the treatment options.

In addition to comparing these two DCM methods, we also report results from a test of an internally developed incentive alignment method, termed Bayesian Truth Serum (BTS). Incentive alignment studies attempt to increase the accuracy of information collected in a survey by rewarding study respondents according to the accuracy of their responses rather than simply for completing the research survey.

While in academic research settings incentive alignment DCM studies have been shown in to deliver greater predictive accuracy than standard DCM methods, commercial market research providers have been stymied in their efforts to migrate the method from academia to the marketplace by the perceived financial and legal hurdles that must be overcome in order to successfully implement an accuracy based cash or product compensation program.

The BTS method that GfK is developing removes these hurdles by implementing a grade-based reward system, whereby respondents receive a letter grade indicating the degree to which they accurately answered key survey questions. The grade, of course, has no cash value and raises no practical or legal concerns. This grading method provides emotional rewards for physicians, as well as a sense of being monitored, recalling their many years in school where accuracy on exams was key to professional prestige and advancement. Initial results indicate that it is particularly effective among physicians, a highly educated population long accustomed to having their skill and knowledge tested and graded.

3. STUDY DESIGN

3.1 Overview

In January of 2017 we conducted an online survey of 400 general practice physicians, half in the US and half in the UK. The topic of the study was the treatment of high BMI Type II diabetes (T2D), and qualifying physicians were limited to those who treat at least 20 high BMI T2D patients. The core of the study was a conjoint exercise where physicians were queried as to their likelihood to prescribe an imagined new treatment for

high BMI T2D patients. It is important to note that in this disease domain it is not uncommon for patients to receive multiple simultaneous treatments. The new treatment profile, referred to as Product X in the survey, was generated from the following attribute profile table:

Attributes	Level 1	Level 2	Level 3	Level 4
Reduction in HbA1c (%)	1.0%	1.2%	1.5%	
Reduction in body weight (kg)	2kg	3kg	4kg	5kg
Impact on systolic blood pressure (mmHg)	3mm Hg	4mm Hg	5mm Hg	
Incidence of hypoglycaemia (%)	1%	2%	3%	4%
Incidence of UTIs / Genital mycotic infections (%)	2%	4%	6%	8%
Flexible doses available	Yes	No		

The set of available current treatments is as follows:

Current Treatments
Metformin
Sulphonylureas
DPP-4 inhibitors
SGLT-2 inhibitors
GLP-1 agonists
Insulin

Within each country the physicians were randomly assigned to complete either the PBC or the ABC version of the survey. Half the respondents in both the PBC and ABC cohorts completed standard versions of the survey (described below) and half were a BTS version of the survey (also described below).

3.2 PBC Survey

The survey administered to physicians in the PBC condition had six main sections:

3.2.1 Patient Record Form and Current Treatment Report

In the first section of the survey physicians completed a patient record form for three of their most recent patients, one each from their populations of low, moderate, and severe patients. For each of these real life patients, physicians recorded numerous details of the patient's clinical and demographic profile. Physicians also reported the treatments currently prescribed for the patient.

3.2.2 Choice Exercise

Physicians then completed a discrete choice exercise where they were asked to select, from a set of three potential treatments, the one they considered to be “most suitable” for

each of the three patients. The profile for each of the treatment options was drawn from the attribute profile table shown above. A “None” option was also provided.

3.2.3 Fixed Profile Calibration Task

Physicians were then shown a sequence of five treatment profiles, each drawn from the attribute profile table shown above, and asked whether they would prescribe the profiled treatment to each of the three patients reported on in the previous sections. The five treatment profiles were presented in sequence, from the one expected by the researchers to be least appealing to physicians (e.g., having low safety and efficacy profiles) to the one expected to be the most appealing to physicians.

In addition, for each of the three patient types (low, medium and high severity) respondents reported the maximum percentage of patients for whom they would prescribe the “best” drug profile instead of their current therapy. We term this the “maximum prescribing percentage.”

3.2.4 Holdout Tasks

Physicians then completed two holdout tasks. In each task they were shown a single product profile and ask to report the percent of their total patient population for whom they would prescribe the profiled treatment if it were available. The holdout task was conducted using a standard allocation format, and respondents were reminded that since patients might receive multiple simultaneous treatments, the percentages reported in the holdout tasks were allowed to sum to greater than 100%.

3.2.5 Physician Peer Question

Physicians were then asked to estimate how likely their peers—physicians similar to them—would prescribe a particular version of the new product profile. An example of the question seen by physicians is shown below. The “Product X” referenced in the question is a version of the new potential product shown in the first holdout task.

So far in this survey we asked you to tell us about your own prescribing behavior. For the following question we would like you to change perspective and think about how other physicians taking this survey will answer. Some of the physicians answering the survey will be similar to you—in age, gender, practice size, etc.—and others will be different. Overall they represent a cross-section of primary care physicians who treat patients with uncontrolled high BMI T2D.

If Product X were available today you indicated you would prescribe it to (**PROG: insert percentage**) of your uncontrolled high BMI T2D patients. Thinking about other physicians completing this survey, please let us know the percent of those physicians who would prescribe Product X to:

		Percent of Other Physicians who would prescribe Product X to each of the following groups of their uncontrolled high BMI T2D patients
1	Less than 10% of their uncontrolled high BMI T2D patients	___%
2	Between 10%–30% of their uncontrolled high BMI T2D patients	___%
3	More than 30% of their uncontrolled high BMI T2D patients	___%

3.2.6 Experience Reports

Finally, physicians were asked a series of questions to gauge the quality of their experience answering the survey.

3.3 ABC Survey

The survey administered to physicians in the ABC condition had five main sections:

3.3.1 Current Prescribing Pattern Report

In the first section of the survey physicians reported, via an allocation format, the percent of patients for whom they prescribed each of the currently available treatments. An example screenshot from the survey is shown below.

% of uncontrolled high BMI T2D patients		
Metformin	<input type="text" value="40"/>	%
Sulphonylureas	<input type="text" value="20"/>	%
DPP-4 inhibitors	<input type="text" value="5"/>	%
SGLT-2 inhibitors	<input type="text" value="25"/>	%
GLP-1 agonists	<input type="text" value="15"/>	%
Insulin	<input type="text" value="20"/>	%
Other (specify) <input type="text"/>	<input type="text"/>	%
Total	125	

3.3.2 Choice Exercise

Physicians then completed a series of 13 choice tasks. In each task they were shown a single product profile and were asked to report the percent of their total patient population for whom they would prescribe the profiled treatment if it were available. These tasks were also conducted using a standard allocation format, and respondents were reminded that since patients might receive multiple simultaneous treatments, the percentages reported in the holdout tasks were allowed to sum to greater than 100%.

3.3.3 Holdout Tasks

Physicians then completed two holdout tasks. Structurally these were identical to the choice exercise tasks. In each task they were shown a single product profile and were asked to report the percent of their total patient population for whom they would prescribe the profiled treatment if it were available. The holdout task was conducted using a standard allocation format, and respondents were reminded that since patients might receive multiple simultaneous treatments, the percentages reported in the holdout tasks were allowed to sum to greater than 100%.

3.3.4 Physician Peer Question

Physicians were then asked to estimate how likely their peers—physicians similar to them—would prescribe a particular version of the new product profile. This was identical to the example shown above.

3.3.5 Experience Reports

Finally, physicians were asked a series of questions to gauge the quality of their experience answering the survey.

3.4 Bayesian Truth Serum

Half the physicians in each the PBC and ABC cohorts completed surveys structured as described above. The other half—those assigned to the BTS cohort—completed surveys identical to those just described with the following three exceptions:

3.4.1 Choice Exercise Text

Before the Choice Exercise respondents in the PBC cohort were shown the following text:

As you complete this section of the survey, please note that in order to increase the validity of this research, we will be grading you on the accuracy of your answers using a five-letter grading system (A, B, C, D, F). This grading method was recently devised by an MIT professor and published in the journal, *Science*. The method rewards you for answering accurately, and the best strategy for receiving a high grade is to carefully consider each question, and answer as accurately as you can.

You will receive your grade in a few weeks after we have collected all responses to this survey.

3.4.2 Holdout Task Text

Before the Holdout Tasks were presented respondents in the PBC cohort were shown the following text:

As with the previous section, your answers here will influence your final grade. And as before, since the grading method rewards you for answering accurately, the best strategy for receiving a high grade is to carefully consider each question, and answer as accurately as you can.

3.4.3 Physician Peer Question Task Text

The same text shown in the holdout task was shown a second time just before the physician peer question was asked.

4. ANALYSIS

The collected data were used to generate two key estimates of prescribing: “prescribing share” and “total prescribing.” Prescribing share is defined as “the percent of patients prescribed a particular treatment” while total prescribing is defined as “the average number of treatments prescribed to each patient.” Below we describe the methods we used to generate these metrics in both the PBC and the ABC domains.

4.1 PBC Data Analysis

Using data collected in the Choice Exercise section of the PBC version of the survey, we estimated conjoint utilities for the varying features of the new product profile via Sawtooth Software CBC/HB using a part-worth estimation procedure. No constraints

were included in the estimation procedure, though whether or not a physician had been exposed to the BTS question was used as a covariate.

We use these utilities to calculate what we term a “threshold utility” for each patient type. This measure represents the level of value a potential treatment must reach in order for physicians to prescribe the treatment. The threshold utility equals the utility sum of the lowest rated product that the physician reports, in the fixed profile calibration task, they would prescribe to their patients. The working assumption is that physicians will not prescribe any product whose utility sum is lower than this threshold utility.

Prescribing shares for any Product X profile are generated in the simulator via the following steps:

1. We first calculate the *preference share* (versus a None option) for the tested Product X profile.
2. We then calculate a *take-up percentage* by comparing the utility sum for the tested Product X profile to the utility level at which a new drug would not be prescribed at all and to the level at which it would be prescribed 100% of the time (as determined via the fixed product profile task). As that utility sum is low in comparison to those levels, we set the take-up percentage to be closer to 0, and as it is high, we set the take-up percentage to be closer to 1.
3. We then calculate the Product X Prescribing Share as follows:
$$\text{Product X Prescribing Share} = \text{Preference Share} * \text{Take-Up Percentage} * \text{Maximum Prescribing Percentage}$$
4. The previous three steps are completed for each of the three patient types and the final Product X prescribing share is calculated as a weighted average of each of the patient types, so the proportion of overweight, moderately obese, and severely obese patients matched the proportion of patients physicians stated they treated at the start of the survey.

The new drug is assumed to steal share from the current treatments in proportionally equal amounts.

4.2 ABC Data Analysis

Using data collected in the Choice Exercise section of the ABC version of the survey, preference shares for each of the 8 treatment options were independently estimated using HB-Reg models. The new product attribute/level indicators served as the independent variables in each model.

Prescribing share for each of the 8 treatments was simply the output of each of the 8 HB-Reg models, expressed as a percentage of total patients. Total prescribing was calculated by summing the output of the 8 HB-Reg models.

4.3 Bayesian Truth Serum

Prescribing share and total prescribing for the BTS cohort were calculated as described above, depending on whether a respondent was assigned to the PBC or the ABC condition.

The letter grade communicated to each respondent was generated by first calculating a BTS score for each respondent using the holdout task and peer question task responses according to the method described in Prelec (2004), and then assigning a letter grade to each score based on the total distribution of BTS scores. The BTS score was not used in the analysis other than as a tool to generate the letter grade.

5. RESULTS

5.1 New Product Prescribing Share

Figure 1 shows the new product prescribing shares that were directly stated in the allocation format holdout tasks in both surveys, as well as the new product prescribing shares predicted by the ABC and PBC models. (In all cases the reported numbers are the average of shares taken across both holdout tasks.)

The PBC model predicts the new product will get 30% share, whereas the ABC model predicts the new product will only get 17% share. Within each cohort the modelled shares closely match the stated responses (PBC—31%; ABC—18%). In both the stated and modelled results, the difference between the two cohorts was significant at the 0.01 level.

Figure 1. Product X prescribing share, both stated and modelled, taken as an average across the two holdout tasks.

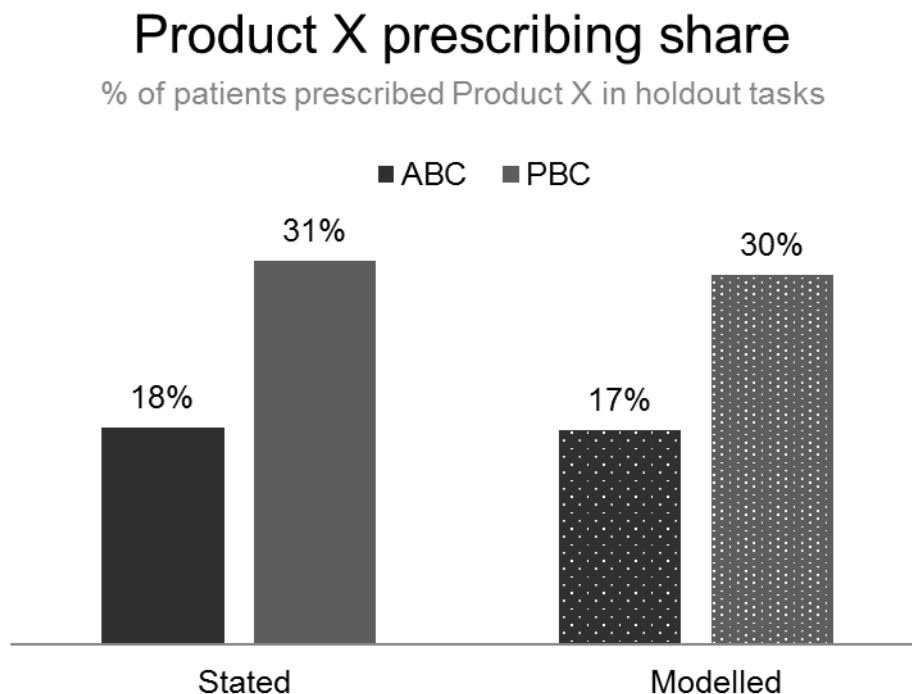
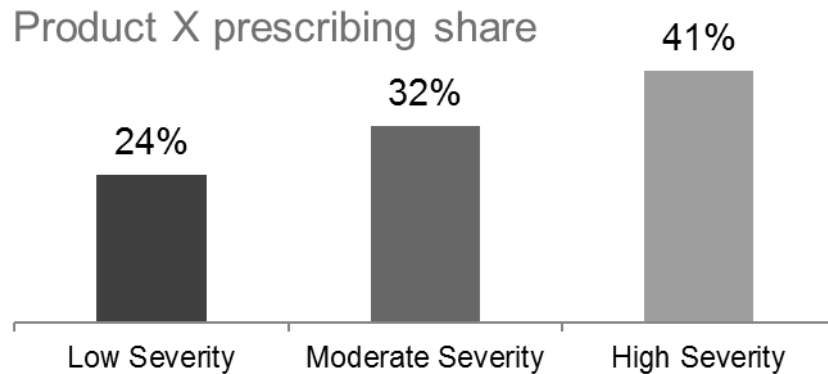


Figure 2 shows the new product prescribing shares predicted by the PBC model for holdout task 1, split by severity of patient. The new product prescribing share is higher the more severe the patient is; low severity = 24%, moderate severity = 32%, high severity = 41%.

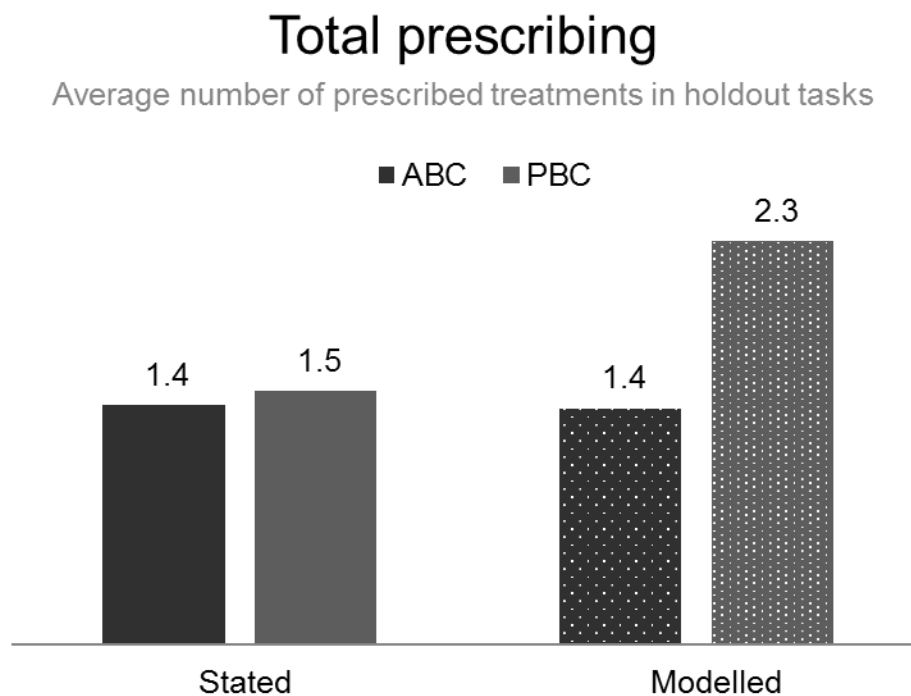
Figure 2. New product prescribing shares, generated by the PBC model for holdout task 1, split by patient severity.



5.2 Total Per Patient Prescribing

Figure 3 shows total per patient prescribing, both stated and modelled, for both the PBC and ABC cohorts. (We define total prescribing as the average number of prescribed treatments per patient. As above, the presented results are an average of responses taken across the two holdout tasks.)

Figure 3. Number of prescribed treatments per patient, taken as an average across the two holdout tasks.



In the stated case, both PBC and ABC cohort physicians report a willingness to prescribe an average of 1.4 treatments per patient. The modelled results, however, show a

different pattern, with the PBC model predicting that patients will receive an average of 2.3 simultaneous treatments, whereas the ABC model predicts only 1.4.

Figure 4 shows the average number of treatments per patient as predicted by the PBC (2.3) and ABC model (1.4) compared with figures from third party sources. By taking an average of five third party sources (see Appendix), we get a figure of 1.7 treatments per patient, which falls between the figures predicted by each model.

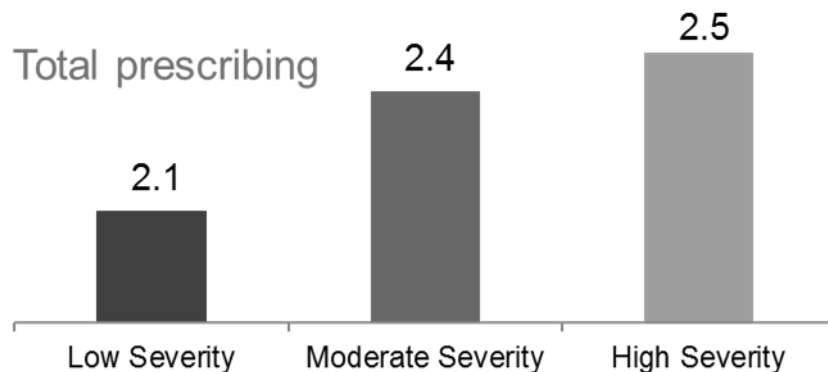
Figure 4. Number of treatments per patient compared to third party data sources.

Data source	Average number of Treatments per patient
ABC Model	1.4
Average of 3 rd party sources	1.7
PBC Model	2.3

(Note that 3rd party data sources reflect treatment levels for all T2D patients, rather than the subset of uncontrolled, high BMI, patients used to generate the ABC and PBC numbers. It may well be that the prescribing rate for this subset of patients is higher than it is for the broader T2D patient population.)

Figure 5 shows the average number of prescribed treatments predicted by the PBC model for holdout task 1, split by severity of patient. The average number of prescribed treatments is higher the more severe the patient is; low severity = 2.1, moderate severity = 2.4, high severity = 2.5.

Figure 5. Average number of prescribed treatments predicted by the PBC model for holdout task 1, split by patient severity.

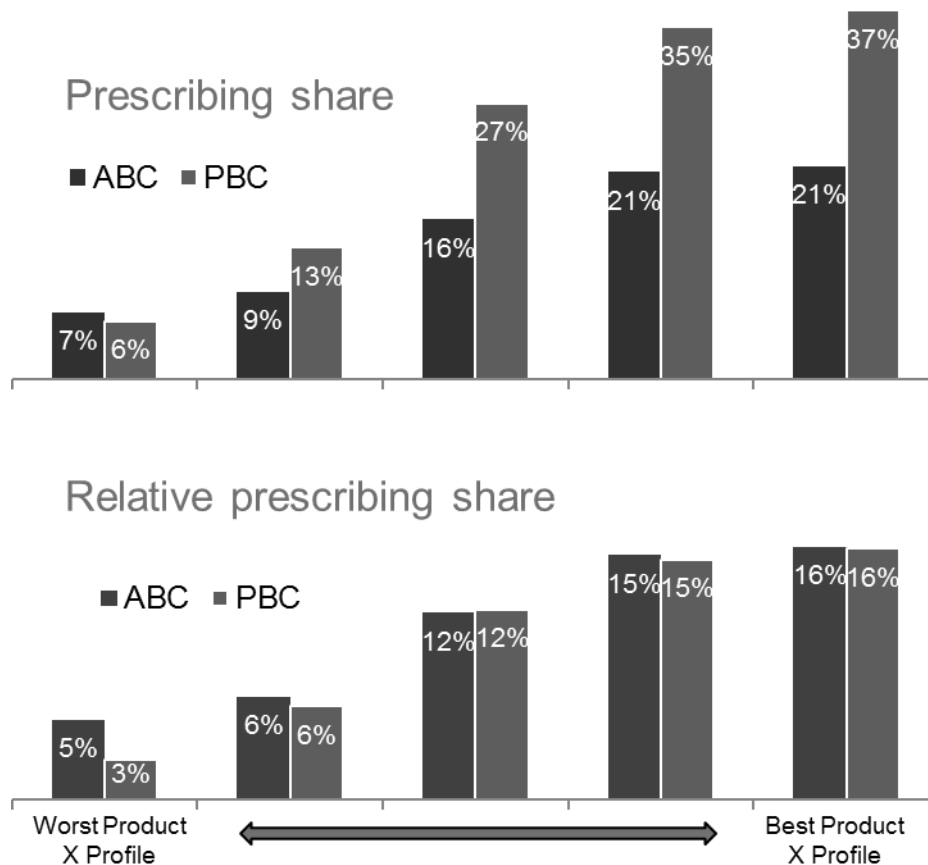


5.3 Relative vs. Absolute Prescribing Share

Figure 6 shows the new product prescribing shares predicted by the ABC and PBC models for 5 Product X profiles ranging from poor to high quality. For the “worst” profile, where each attribute is set to the “worst” level, the shares predicted by the ABC and PBC models, 7% and 6% respectively, are similar. However, for the “best” profile, where each attribute is set to the “best” level, the 37% share predicted by the PBC model is much higher than the 21% share predicted by the ABC model. Hence, changes in the new product profile causes a larger change in share in the PBC model compared with the ABC model.

However, when the shares are rescaled so that the total share across all products in the model sums to 100% (previously this figure was greater than 100% as detailed above), the resulting “relative prescribing shares” are comparable.

Figure 6. Simulated preference shares for 5 Product X profiles ranging from poor to high quality.



5.4 Sensitivity Analysis

We measured each model’s sensitivity by first setting each attribute to its “worst” level then recording the uplift in new product prescribing share obtained as each attribute

is independently changed to its “best” level. The uplifts are expressed as percent change in prescribing level.

The uplifts are higher across all attributes in the PBC model compared to the ABC model. For example, for the “most important” attribute, “Reduction in HbA1c (%)” the uplift is 10.9% in the PBC model compared to 2.4% in the ABC model.

Figure 7. Absolute and relative attribute sensitivity.

Attributes	ABC	PBC	ABC	PBC
	Sensitivity		Relative Sensitivity	
Reduction in HbA1c (%)	2.4	10.9	30%	30%
Reduction in body weight (kg)	2.0	9.5	25%	26%
Incidence of UTIs / Genital mycotic infections (%)	2.0	8.7	25%	24%
Incidence of hypoglycaemia (%)	0.6	3.3	8%	9%
Flexible doses available	0.1	2.0	1%	6%
Impact on systolic blood pressure (mmHg)	0.9	1.5	11%	4%

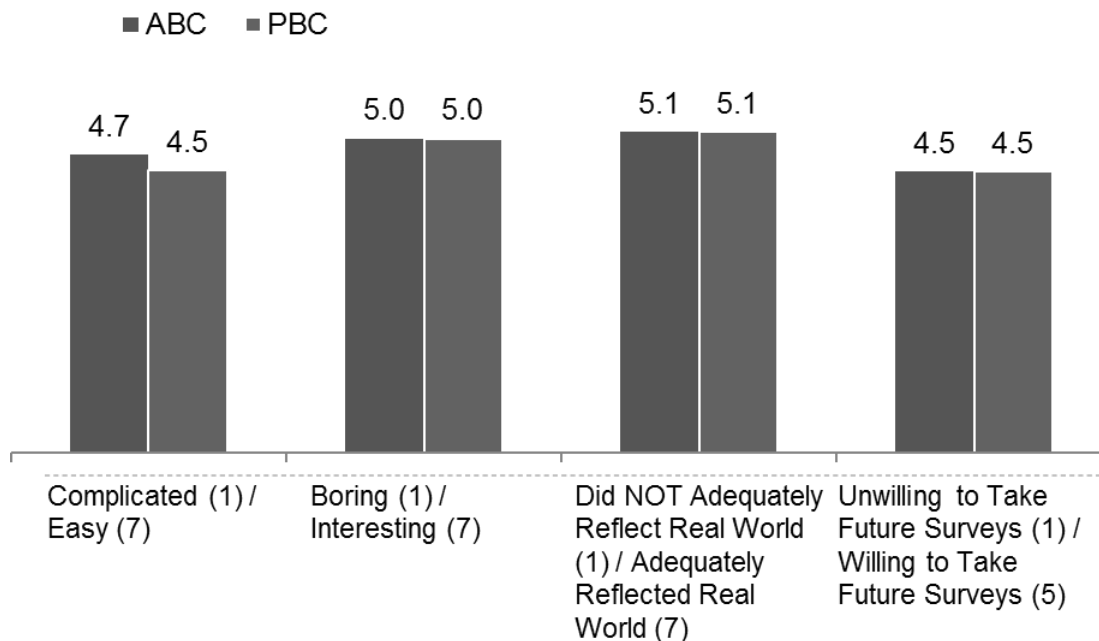
However, when the uplifts within each model are rescaled relative to each other, so that the uplifts in each model sums to 100%, the resulting “relative sensitivities” of each attribute are again comparable. The “relative sensitivity” for the four “most important” attributes is almost identical.

5.5 User Experience Comparison

Three user experience questions were included at the end of each survey to understand differences in user experience across the different methodologies. The questions were asked on a 7-point anchored scale. An additional question about willingness to answer a similar survey again was also asked on a 5-point anchored scale.

Figure 8 shows that there was little difference between the responses of the physicians completing the two surveys. No significant differences were found at the 0.05 level.

Figure 8. User Experience Ratings



5.6 Bayesian Truth Serum

As we discuss below, the previous analysis strongly suggests that a significant number of respondents did not correctly understand (or follow) the survey instructions. For example, in the PBC cohort it appears that respondents may not have reported data from their most recent patients, but rather for a subset of those recent patients whose severity makes them top of mind. It is also likely that a certain percentage of physicians assumed that their responses to the allocation format holdout tasks were required to sum to 100% despite receiving explicit instructions that this was not the case. Essentially, these physicians acted as if they were being asked to report share of total prescriptions accounted for by each treatment, rather than share of patients receiving each treatment.

This incorrect reading of the allocation question—if it occurs—would mask much of the effect of the BTS methodology, which is primarily expected to counter the survey induced bias of physicians to overreport their potential prescribing. Therefore, in the subsequent analysis we excluded those respondents whose total per patient prescribing equaled 1 in both holdout tasks. In other words, we limit our BTS analysis to those physicians who report prescribing, on average, more than 1 prescription per patient. Note that this results in reported prescribing numbers that differ from those reported in the previous section of the report.

Figure 9 shows that the PBC cohort physicians in the BTS case reported a lower Product X prescribing share than did those in the non-BTS case (difference significant at the 0.05 level), while no BTS effect was seen in the ABC cohort. Figure 10 shows a similar pattern when Total prescribing is used as the outcome measure.

Figure 9. Effect of BTS on Product X prescribing share by PBC and ABC cohorts.

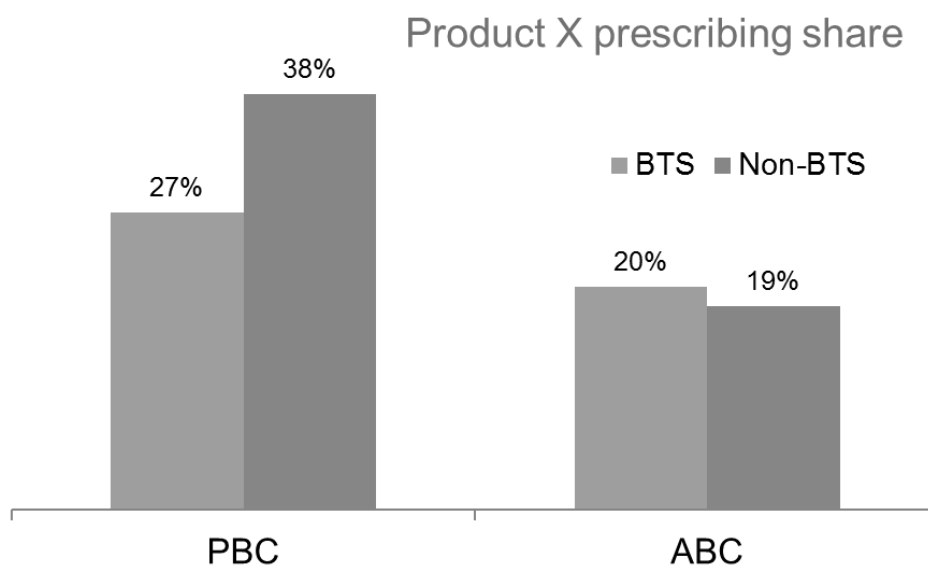


Figure 10. Effect of BTS on Total prescribing, by PBC and ABC cohorts.

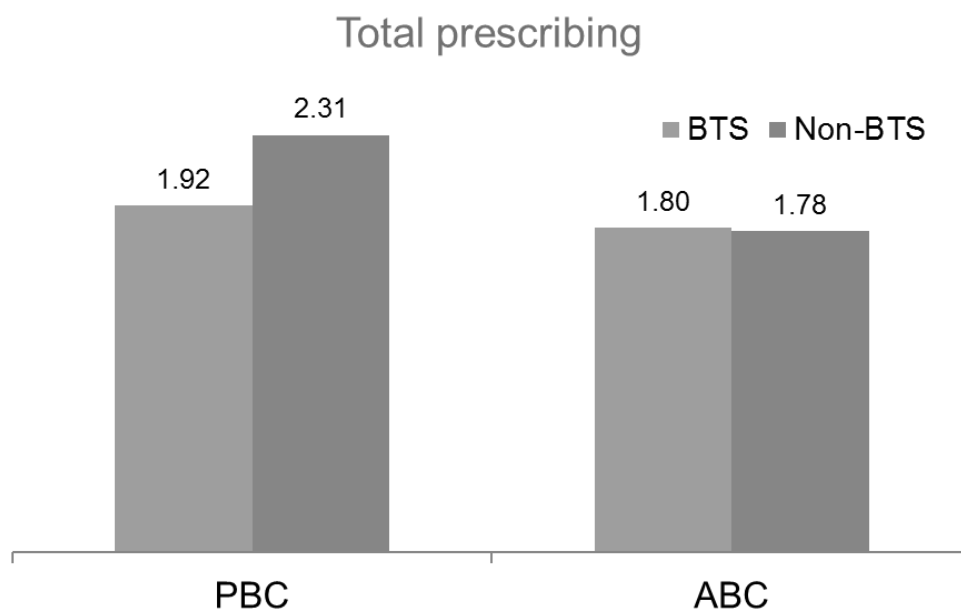
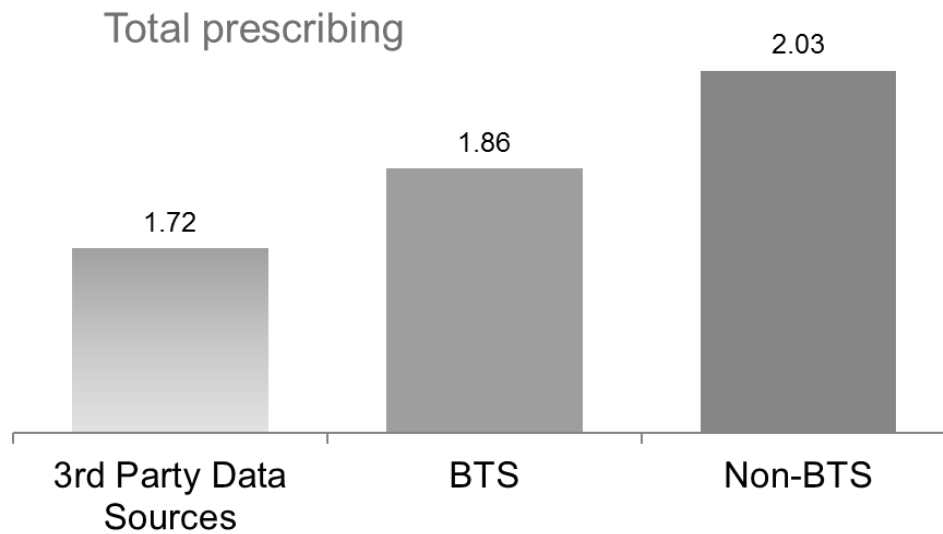


Figure 11 shows that BTS total prescribing estimates more closely match 3rd Party estimates than do non-BTS estimates. BTS and Non-BTS figures include both ABC and PBC cohorts. Numbers reported for 3rd Party data sources represent an average of the available sources.

Figure 11. Comparison of total prescribing as reported by 3rd party data sources to both the BTS and non-BTS cases.



6. DISCUSSION

The study shows the profound impact that question format can have on physician estimates of their own prescribing behaviour and, subsequently, on researcher estimates of preference share and market impact. Furthermore, the results emphasize that the standard tools available to researchers to control for model accuracy—comparing modelled results to holdout tasks—are best described as internal consistency checks rather than as methods capable of gauging the predictive accuracy of the survey. The findings of this study well exhibit this point, as the PBC and ABC models generate dramatically different preference share estimates, yet both models accurately predict physician responses to the holdout tasks.

6.1 The Patient Based Model Potentially Overestimates Share

Compared to the ABC model, the PBC model predicts higher prescribing shares for the preferred profiles of the new product and a higher average number of prescribed treatments per patient. One hypothesis is that the PBC model overestimates shares.

In the PBC survey we asked physicians to think of their three most recent patients who had type 2 diabetes, were uncontrolled on their current treatment, and had a high BMI. The hope underlying this instruction is that it would result in a representative sample of the actual patient population. It may well be that the instruction actually led physicians to select a patient sample skewed to more involved patients, with more extreme characteristics than average.

This might be the case because extreme patients are naturally more memorable and likely to have seen the physician a greater number of times. Therefore, our sample of patients may be more extreme than what is representative of this subpopulation of T2D patients. The results show that both new product prescribing share and the average

number of prescribed treatments per patient both increase as the severity of their obesity increases, which suggests, if such a bias is present, it would lead to the PBC model overestimating shares.

However, this is controlled to some extent by the fact we had quotas on how many patients fell into the low, moderate and severely obese sub-populations and weighted the sample so the proportion of patients in each group matched the proportions physicians stated they treated earlier in the questionnaire. However, patients could still have been more severe than average within each severity band.

The fact that within the lower severity group the new product prescribing shares (24%) and the average number of prescribed treatments per patient (2.1) is still much higher than predicted by the ABC model, 17% and 1.4 respectively, suggests even if this bias exists it is not enough to fully explain the share differences between the models. Of course, there could be factors other than obesity that make a patient more extreme in their characteristics and therefore more likely to receive a greater number of treatments.

6.2 The Allocation Based Model Potentially Underestimates Share

Conversely, there is evidence to suggest that the ABC model underestimates share. Amongst of the physicians completing the PBC survey, 57% stated their patients would be on a monotherapy in the allocation format of the question, whereas the same physicians stated only 26% of patients would be on a monotherapy via the multi-punch question in the patient record form. 57% monotherapy feels intuitively much too high in a therapy area where we know patients are commonly on multiple treatments simultaneously. Perhaps some respondents thought the allocation had to sum to 100%, or they find it cognitively difficult to make numbers sum to more than that and are attracted by the nice round number 100 represents. Having too many physicians only allocating a total share sum of 100% would naturally lead to the ABC model underestimating the average number of prescribed treatments per patient.

6.3 Difference in Stated Responses to Holdout Tasks

It is curious that the stated responses to the exact same two holdout tasks is so markedly different, with physicians completing the PBC survey stating a new product prescribing share of 31% compared with 18% given by physicians completing the ABC survey. It appears the survey questions the physician answers prior to completing the holdout tasks has a large influence on how much share they allocate to the new product.

One hypothesis is that the patient record form engages physicians to think about their patients in much more detail and they therefore identify a greater proportion of patients that the new drug could be prescribed to. The lack of patient detail in the allocation survey means physicians miss certain patients that would be good candidates to receive the new drug.

An alternative hypothesis is that there is an exposure bias towards the new product in the PBC survey. Physicians completing the PBC conjoint exercise are exposed to more profiles of the new product as three new products are shown per task versus only one per task in the ABC. This greater exposure may lead to the physician being more likely to allocate more share to the new product. Also, since these three new product profiles are

only evaluated versus a “none” option it is likely that physicians will pick a new product with some regularity, which creates an affinity for the new product, stronger than that gained from the allocation conjoint exercise, where share is allocated to the new product in a competitive context.

6.4 Impact of BTS on Physician Responses

The BTS method mitigates DCM response biases, and generates prescribing estimates that more closely match those reported by validated third party data sources. It also appears to minimize some of the response biases inherent in survey research.

6.5 Recommendations

Researchers should pay careful attention to how we ask the prescribing questions in our survey, as the question format can have significant impact on survey responses. Instructions given to the respondents must be clear in order to minimise misinterpretation, and testing should be done before a survey is fielded in order to gauge the degree to which respondents are accurately understanding the questions. In the PBC method, perhaps detailed guidelines should be given on how to select sample patients. In the ABC method, perhaps respondents whose allocation sums to 100% could be reminded that the allocation is allowed, and even expected, to sum to more than 100%.

Other recommendations:

- When using the ABC methodology, ask the allocation question by patient type to ensure that different types are well represented.
- Conduct more research to improve PBC methodology and add flexibility for further calibration of the prescribing shares estimates.
- Compare two methodologies in a monotherapy disease domain where high quality 3rd party prescribing data exist.



James Pitcher



Tatiana Koudinova



Daniel Rosen

REFERENCES

- Ding, M. (2007) An Incentive-Aligned Method for Conjoint Analysis. *Journal of Marketing Research*, 44 (May), 214–223.
- Ding, M., Grewal, R., & Liechty, J. (2005) Incentive-Aligned Conjoint Analysis. *Journal of Marketing Research*, 42 (February), 67–82.

- Ding, M., & Hauser, J.R., Dong, S., Dzyabura, D., Yang, Z., Su C., & Gaskin, S. (2011) A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing Research*, 27, 25–32.
- Dong, S., Ding, M., & Huber, J. (2010) A simple mechanism to incentive-align conjoint experiments. *International Journal of Research in Marketing Research*, 27, 25–32.
- Higgins V., Piercy J., Roughley A., Milligan G., Leith A., Siddall J., & Benford M. Trends in medication use in patients with type 2 diabetes mellitus: a long-term view of real-world treatment between 2000 and 2015. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*. Volume 2016:9 Pages 371–380
- Miller, K. M., Hofstetter, R., Krohmer, H., & Zhang, Z.J. (2011) Unstructured Direct Elicitation of Decision Rules. *Journal of Marketing Research*, 48 (February), 116–127.
- Prelec, D., & Weaver, R. (2010) Creating Truth-telling Incentives with the Bayesian Truth Serum. 2010 International Conference on Modelling and Simulation in Engineering, Economics and Management.
- Prelec, D. (2004) A Bayesian Truth Serum for Subjective Data. *Science*, 306/5695 (October 15), 462–466.
- Roper Diabetes Patient Market Study. GfK Healthcare, 2015
- Sharma M., Nazareth I., Petersen I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. *BMJ Open* 2016;6:e010210. doi:10.1136/bmjopen-2015-010210

A META-ANALYSIS ON THREE DISTINCT METHODS USED IN MEASURING VARIABILITY OF UTILITIES AND PREFERENCE SHARES WITHIN THE HIERARCHICAL BAYESIAN MODEL

JACOB NELSON
EDWARD PAUL JOHNSON
BRENT FULLER

RESEARCH NOW—SURVEY SAMPLING INTERNATIONAL

INTRODUCTION

Market researchers employing hierarchical Bayesian (HB) analysis on conjoint and MaxDiff data have often employed shortcut methods to compute share predictions from estimated models. Specifically, many researchers rely solely on the mean of the posterior draws of the lower-level part-worth parameters (“betas”) in their calculation of preference share. Sawtooth Software, in particular, uses this method in its standard utility and preference output. Those using the Sawtooth Software method estimate the variance around the utility means and preference shares by treating each point estimate as a measured value and using $p(1 - p)/\sqrt{n}$ or s/\sqrt{n} , where p is the aggregate preference share estimate, s is the standard deviation of individual point estimates in the lower-level model, and n is the sample size (Sawtooth Software Inc., 2009).

Other researchers use all the posterior draws from the lower-level model (Chapman & Feit, 2015). Those using this method calculate preference share for each respondent for each of the converged beta draws, and average those shares within each iteration. With the preference share estimates for each draw, they then take the 95% credible interval (2.5% and 97.5% percentiles) and use this to estimate the variance around their estimates.

Lastly, some researchers emphasize using the posterior draws for the upper-level model parameters instead. These researchers use the mean and covariance matrix for the upper level to calculate utilities and simulate preference shares. They then take the average of the parameters of interest in each posterior draw and calculate a 95% credible interval across the converged draws. This method will focus just on the overall population parameters of interest, and if specific subset analysis is needed use covariates to examine how subsets of the data differ from the overall population (Allenby et al., 2014; Lee, 2016; Kurtz & Binner 2016). Relying on the upper-level posterior distribution in this manner requires the researcher to rerun the model every time with the appropriate covariates when a new subgroup of interest comes up, which may not be acceptable in many research situations.

To summarize, there are three popular methods to simulate using HB conjoint data:

1. Estimate uncertainty with the formula $p(1 - p)/\sqrt{n}$ or s/\sqrt{n} .
2. Use the posterior draws from the upper-level model (population mean part-worths) to estimate uncertainty.
3. Use the posterior draws from the lower-level model (individual-level part-worths) to estimate uncertainty.

Our goal with this study is to compare these three methods.

We analyze the results from all three methods for 50 conjoint or MaxDiff research projects conducted by Survey Sampling International in 2017. We compare the impacts of these methods across both types of projects. In comparing each of these three simulation methods, we focus on the mean and interval width of each part-worth utility and the mean and interval width of the preference share given a specific market condition.

MOTIVATING EXAMPLE

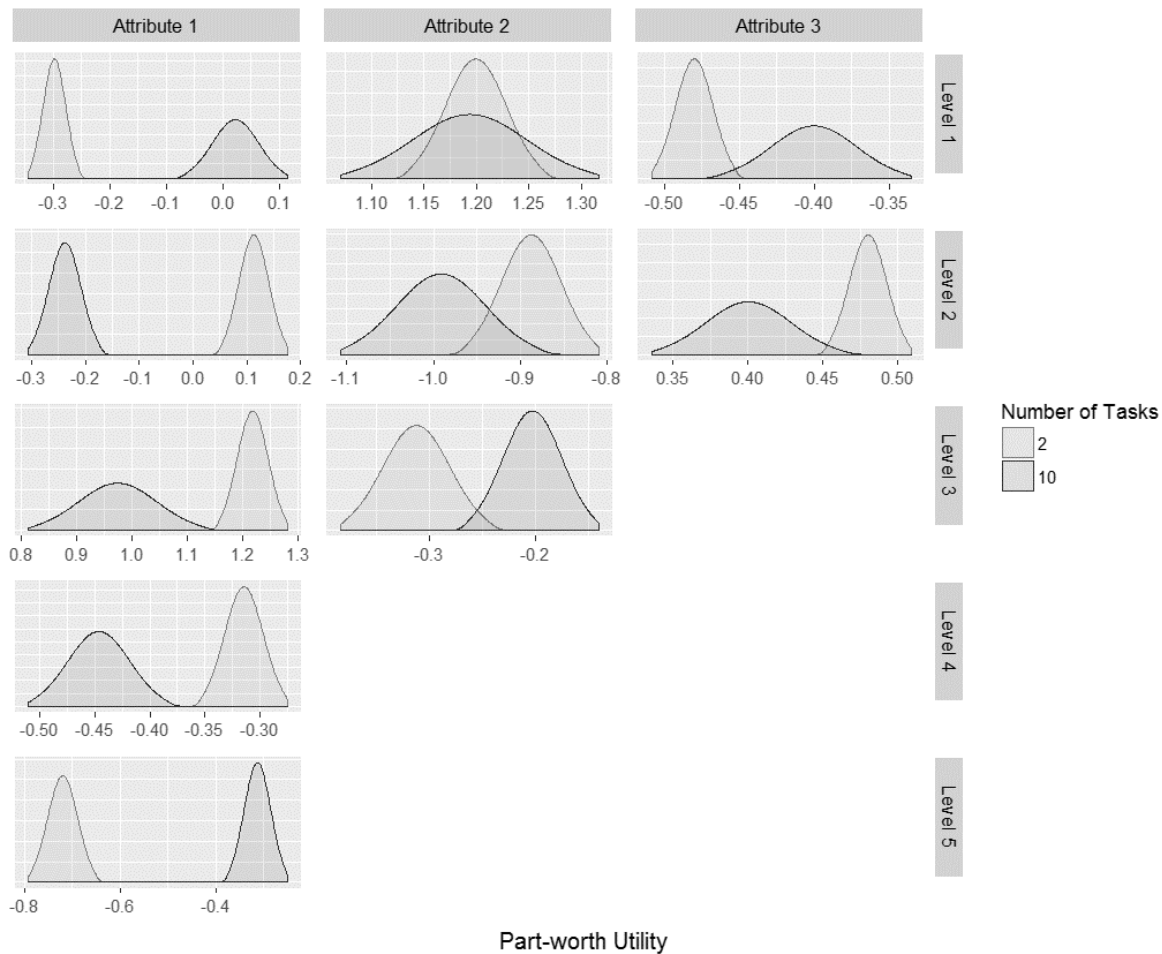
Greg Allenby revolutionized conjoint analysis in market research by introducing the hierarchical Bayesian (HB) model (Allenby & Ginter, 1995). The HB model captures respondent-level heterogeneity in product preferences extremely well. This advantage comes by allowing each respondent to have his or her own preference parameters. Everyone's part-worth utility (point estimate *and uncertainty*) is captured when simulating from the posterior distribution (Allenby & Rossi, 2003). It also contains an upper-level posterior distribution which could be used to simulate from the entire population rather than just those composed in the sample (Allenby et al., 2014).

While all these elements of the model are available for use in a market simulator, many researchers instead employ a shortcut suggested by Sawtooth Software to simplify the computation in a market simulator. Researchers instead take the converged iterations of the HB model and average the part-worth estimates for each respondent (can be exported in the utilities.csv file) as found in the default market simulation calculations for Sawtooth Software products. This point estimate method retains the respondent-level heterogeneity in the average part-worth estimates, but does not consider the degree of uncertainty around those estimates.

Ignoring the individual uncertainty in this way can lead to some non-intuitive results. In general, more choice tasks should increase our certainty about the part-worth utilities (decrease the variance of the aggregate estimates) as more information should decrease uncertainty. However, using the point estimate method, which ignores the uncertainty around a single individual's part-worth utility, you see the opposite result. For example, we consider a conjoint data set with 709 individuals completing 10 choice tasks containing 3 attributes with 2–5 levels each. With the data, we estimate two HB models; one that uses only 2 of the 10 available choice tasks, while the other uses all 10. We take the point-estimates from both HB models and graph normal distributions for the mean part-worths of each parameter. Figure 1 demonstrates how the point estimate method predicts that collecting more information per respondent *decreases* the amount of certainty we have around the aggregate part-worth utility estimates! How can more data lead to more uncertainty?

Figure 1

Method of Estimating Uncertainty: Point Estimate



Part-worth Utility

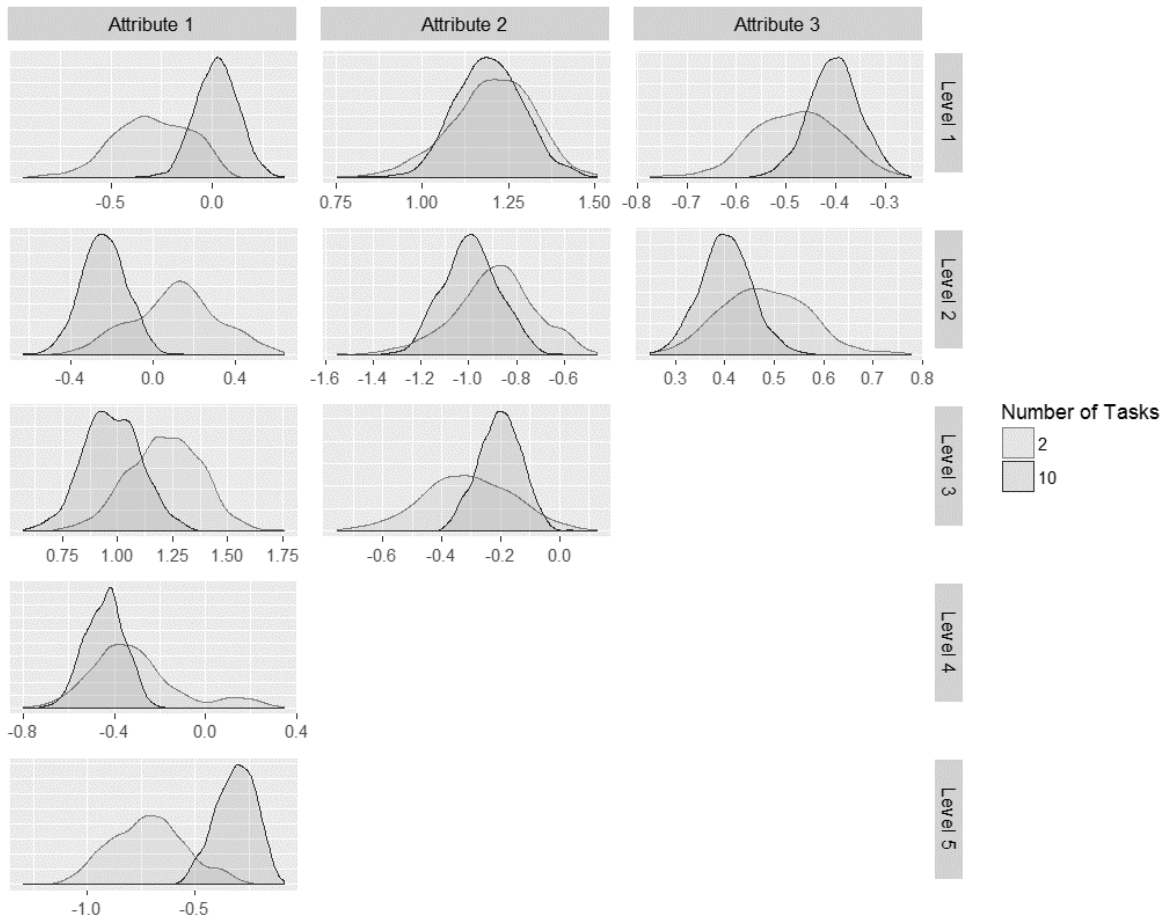
This figure shows how the point estimate method creates a counter-intuitive results where more data (10 tasks versus 2 tasks) leads to less certainty around the aggregate part-worth utilities (wider distribution).

The underlying cause of this paradox from the point estimate method is Bayesian shrinkage. Bayesian shrinkage is a result of borrowing information from the overall group of respondents when there is not enough information from a single respondent (Morris, 1983). The lower-level part-worth estimates in a Bayesian model will tend to shrink towards the overall mean when there is less data on each respondent. Because the point estimate method doesn't consider the individual-level certainty, it will erroneously view the Bayesian shrinkage as actually reducing the amount of heterogeneity among the individuals.

The upper-level or lower-level posterior distributions do account for the variability of the individual level and thus show the appropriate narrower confidence intervals when more data is collected. Figures 2 and 3 use the same data set put forth in Figure 1 to demonstrate how more tasks should decrease the uncertainty rather than increase it. In both figures, the uncertainty about the utility of each feature is predicted to be smaller when there are more observed choices, that is, the darker distributions are narrower than the lighter ones.

Figure 2

Method of Estimating Uncertainty: Upper Level



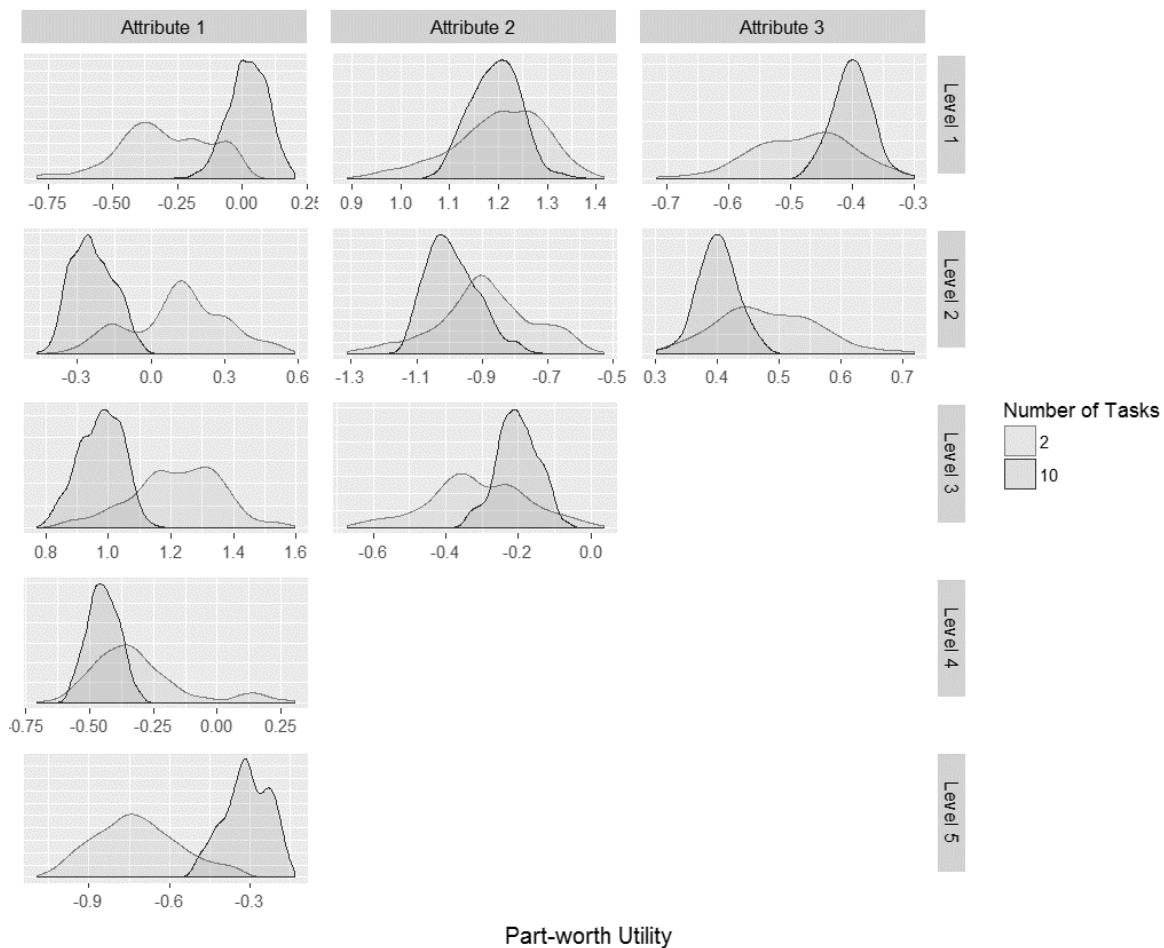
Part-worth Utility

This figure shows how the upper level method creates an intuitive results where more data (10 tasks versus 2 tasks) leads to more certainty around the aggregate part-worth utilities (wider distribution).

Luckily, Sawtooth Software is also capable of outputting either the posterior draws for the upper-level posterior distribution parameters (the `alpha.csv` and `covariance.csv` files) or the lower level posterior distribution parameters (`draws.csv`) by iteration for those who would like to use either of these methods to construct a market simulator. While these market simulators might be more difficult to construct in Excel, they are certainly options for market researchers to use, given the computing power available today.

Figure 3

Method of Estimating Uncertainty: Lower Level



This figure shows how the lower level method creates an intuitive results where more data (10 tasks versus 2 tasks) leads to more certainty around the aggregate part-worth utilities (wider distribution).

RESEARCH DESIGN

Our analysis includes 50 HB choice models run by SSI in the marketing research field across multiple methodologies, industry sectors and model characteristics. We include 26 conjoint projects, 22 of which are CBC and 4 of which are ACBC. 12 of the 26 have alternative-specific designs. We also include 24 MaxDiff studies, with the number of items ranging between 13 and 42 depending on the project. Sample sizes across these 50 projects varied widely, between 50 and 6,800 respondents. Number of tasks shown ranges from 6 to 15, and number of attributes ranges from 3-20.

All models were estimated with Sawtooth Software products, either with Lighthouse Studio or CBC/HB. Convergence was monitored by viewing trace plots and enough iterations were included in each model to ensure convergence. At least 1,000 draws were saved after the burn-in period on each model using a skip factor of 5 or more for saving random draws. While many researchers use covariates, especially when using the upper level, SSI does not have the industry

vertical expertise to always incorporate meaningful covariates. As such, no covariates were used in any model, each using default settings for the prior alpha and covariance matrix. Constraints were sometimes applied (e.g., negative price constraints) on select projects based on original project specifications.

We explored the posterior distributions of each model and compared the impacts of Bayesian versus frequentist methods of assessing uncertainty using three methods: 1) Using point estimates based on frequentist statistical methods, 2) using the HB lower level posterior distribution and 3) using the HB upper level posterior distribution.

Point Estimate Method

The point estimate (frequentist) method averages the part-worth utilities from the lower-level posterior distribution into a single measurement of preference. Share estimates apply the logit rule to the summed part-worth utilities of each configuration. Uncertainty is then measured by calculating confidence intervals for means using $\bar{x} \pm 1.96(s/\sqrt{n})$ and proportions using $1.96\sqrt{\hat{p}(1 - \hat{p})/n}$ as if preference was a measured variable (such as weight or height of a person) and not a model prediction.

HB Lower-Level Posterior Distribution Method

The lower-level posterior distribution method directly uses the beta draws, which are the individual-level part-worths for each respondent. Average utilities are generated and share of preference is calculated using the logit rule across all saved posterior draws for each respondent. Uncertainty is measured by taking the 2.5% and 97.5% percentiles of the saved draws.

HB Upper Level Posterior Distribution Method

The upper-level posterior distribution method draws from the alpha and covariance files simulating “synthetic respondents” for each draw. From these synthetic respondents, we calculated part-worth utilities as well as the share of preference for each product in the simulated preference using the logit rule. To accomplish this, the upper model covariance matrix produced under Sawtooth Software dummy coding option needs to be used and converted to a symmetric matrix (see Appendix A for R code). With that we then simulate from the multivariate normal distribution to generate the synthetic respondents. Note that for the MaxDiff projects, dummy coded alphas and the covariance matrix output can be obtained through CBC/HB rather than Lighthouse Studio. Once this simulation was complete we then averaged these metrics across multiple (up to 5,000) draws or synthetic respondents repeated for every iteration in the upper level model and then take the 2.5% and 97.5% percentiles for the credible interval.

In each study and for each method we measured the following:

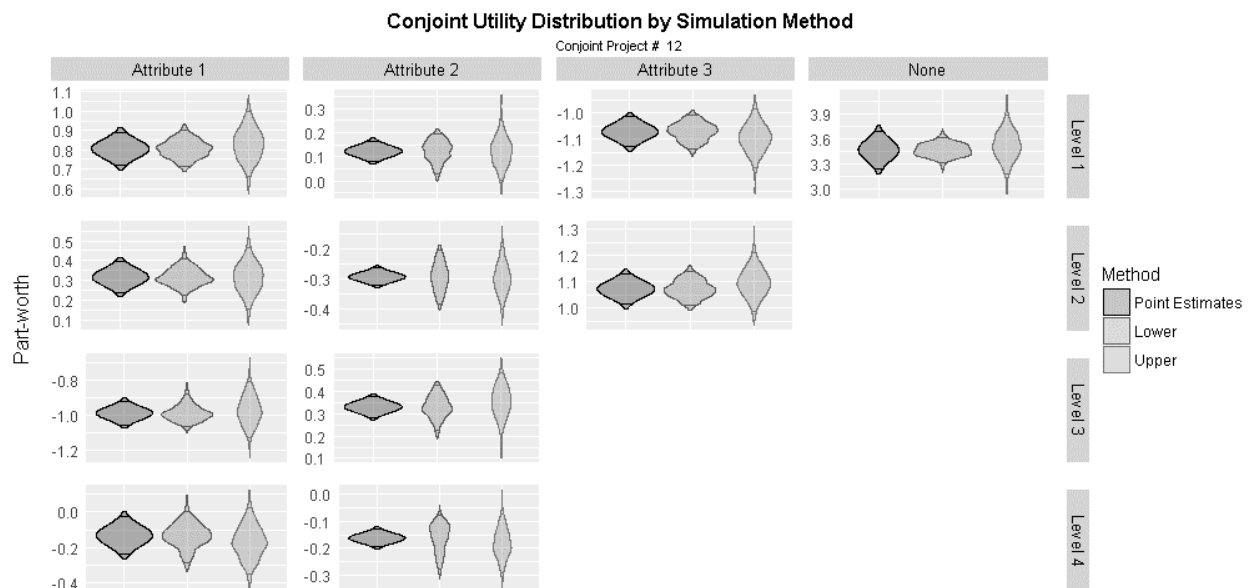
1. Average utilities for each parameter
2. Average utility interval width for each parameter
3. Average distances from the point estimate for Preference Share from 10 configurations
4. Average Preference Share interval widths averaged across 10 configurations

RESULTS

Conjoint Results—Utilities

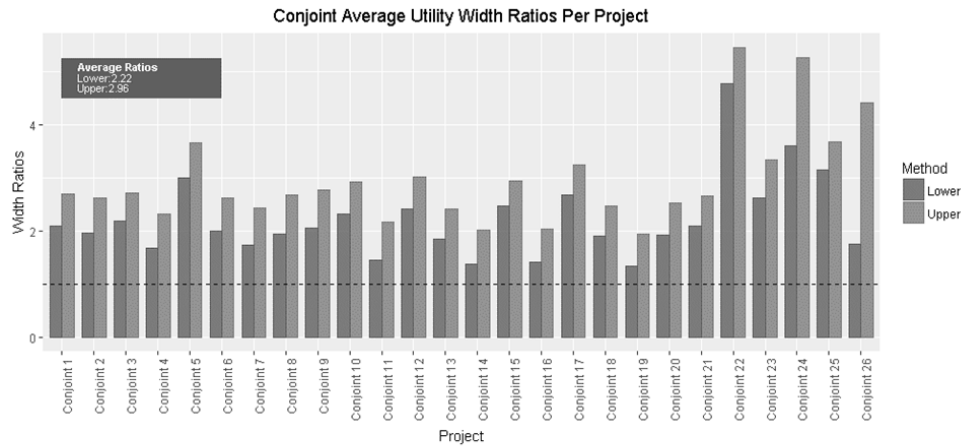
Figure 4 demonstrates an example of one of the CBC project's part-worth distribution, and similar results were yielded from other conjoint projects. Each attribute level's mean part-worth utility distribution is shown in a violin plot by estimation method. Each method has a mean center in the same place for each level, which is to be expected, as the point estimate *is* the mean of the lower level posterior distribution, and the upper level posterior distribution should converge to the same part-worth as well. However, the width of each distribution is quite different between methods. This reflects different estimates of how much certainty we have about the part-worths. As you can see from the figure, the uncertainty in the utilities is greatest for the upper level model, somewhat smaller for the lower-level model method and smallest for the formula method. This means that the simple formula method is often understating the uncertainty in part-worths.

Figure 4



If we compute a ratio between the widths of 95% credible and confidence intervals between methods, we learn that the lower level posterior distribution's credible interval is, on average, 2.2 times larger than the point estimate's confidence interval, and the upper level posterior distribution method's credible interval is, on average, 2.9 times larger. Figure 5 shows these ratios per project in a bar chart. In all 26 of these conjoint projects, the point estimate method underestimated the confidence interval compared to these other two methods, sometimes by as much as nearly 5 times the point estimate's distribution width. While the underestimated variance may not alarm some researchers (many who do not do significance testing on the conjoint results anyways), it can lead to substantially different share of preference estimates as we show in the next section.

Figure 5

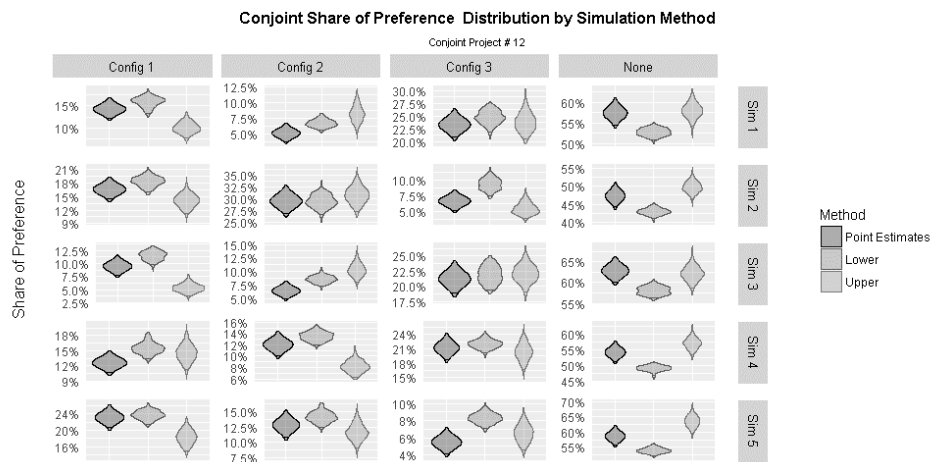


This figure represents the average "width" ratios from frequentist and posterior distributions of part-worths from multiple conjoint projects. For every model parameter in every project, we took 95% and 5% quantiles and measured the distances between them, in frequentist and posterior distributions. A ratio of 1 (represented by the dotted line) signifies the point where the width (uncertainty) of the frequentist and posterior (upper, lower) distributions are the same. We can see that frequentist distribution tended to overestimate the width of the posterior distributions, sometimes by a factor of 3 or 4.

Conjoint Results—Shares of Preference

Figure 6 shows share of preference distributions for the three methods, split by configuration and then simulation scenario. The scenarios shown below are typical for other simulations and among other projects. The distributions of shares of preferences show not only differences in distribution widths, but also in **differences in where those distributions are centered** (their means). Moreover, these differences are often practically and statistically significant when tested with a chi-squared test. Table 1 shows the average percent of simulations that have significant differences between product share of preference estimates.

Figure 6

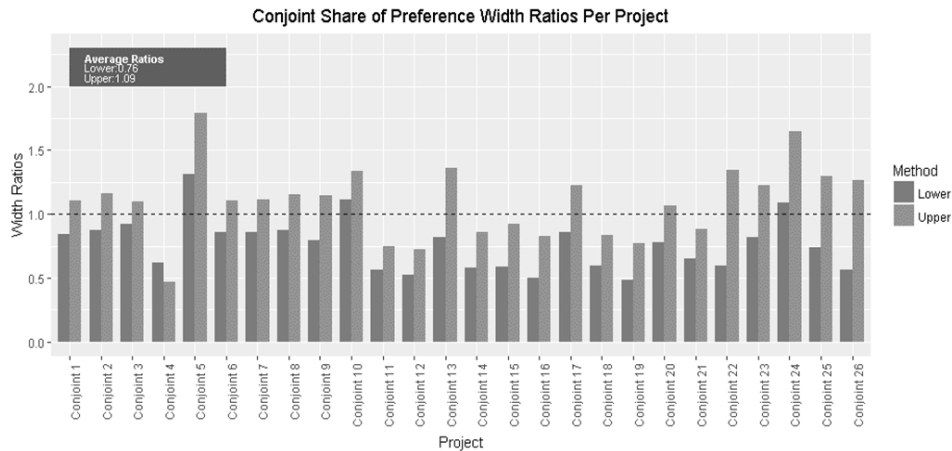


This figure shows share of preference distributions from frequentist and posterior (upper, lower) distributions from an example HB model derived from a choice based conjoint exercise. Each column of grids represent a product configuration, and each row represent a new simulation. The last column represents the "None" choice. We can observe from the figure that distributions of share of preference are not centered at the same place and sometimes do not even overlap.

Table 1

	Lower	Upper
Point Estimate	12.1%	44.3%
Upper	51.7%	

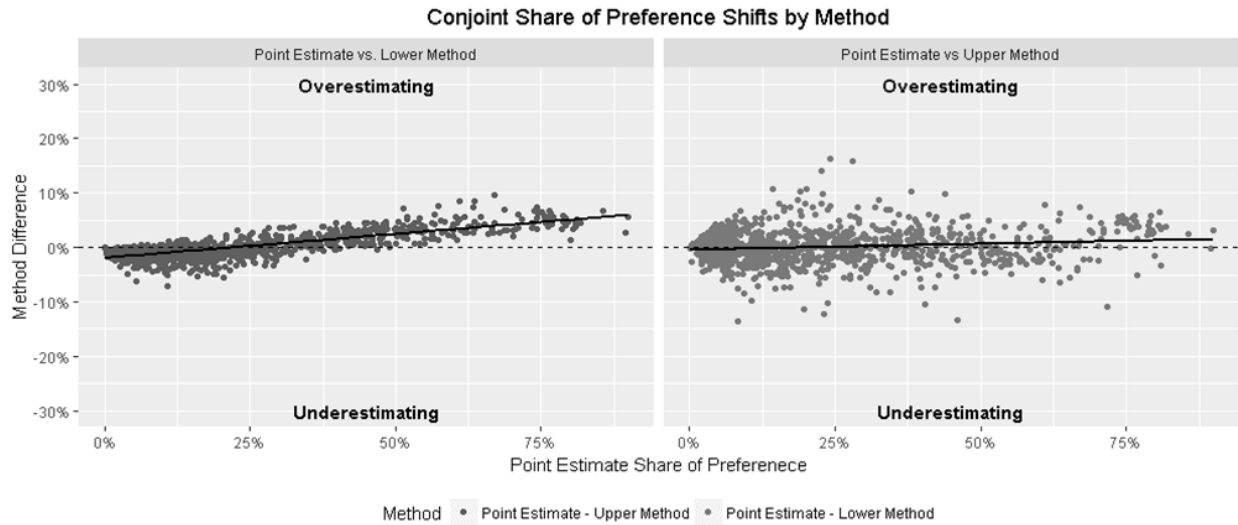
Figure 7 shows again the project to project average distribution width ratios, this time for share of preference. Differences in these widths varied widely from project to project, but showed that on average, the point estimate method tended to overestimate the spread compared to the lower level posterior distribution method, with a mean ratio of 0.76, and slightly underestimate the spread compared to the upper level method, with a mean ratio of 1.09.

Figure 7

This figure represents the average "width" ratios from frequentist and posterior distributions of share of preference from multiple conjoint projects. For every model parameter in every project, we took 95% and 5% quantiles and measured the distances between them, in frequentist and posterior distributions. A ratio of 1 (represented by the dotted line) signifies the point where the width (uncertainty) of the frequentist and posterior (upper, lower) distributions are the same. We can see that the frequentist distribution widths are very different from posterior distributions, at times overestimating and at others underestimating uncertainty

During our research, we noticed a pattern among particularly the lower-level posterior distribution method share of preference compared to the point-estimate method. The mean point-estimate share of preference tended to underestimate the low mean share estimates from the lower level method, and vice-versa. This was not as clearly observed when comparing the point estimate shares to those resulting from the upper-level method. The clear linear trend in the lower-level method from Figure 8 demonstrates this observation. Shares of preference estimates tend to be more flat in the lower level model (i.e., in a 4-product simulation, shares are more likely to be closer to 25%). It should be researched in the future whether this is the result is capturing more respondent noise, indifference, or both.

Figure 8

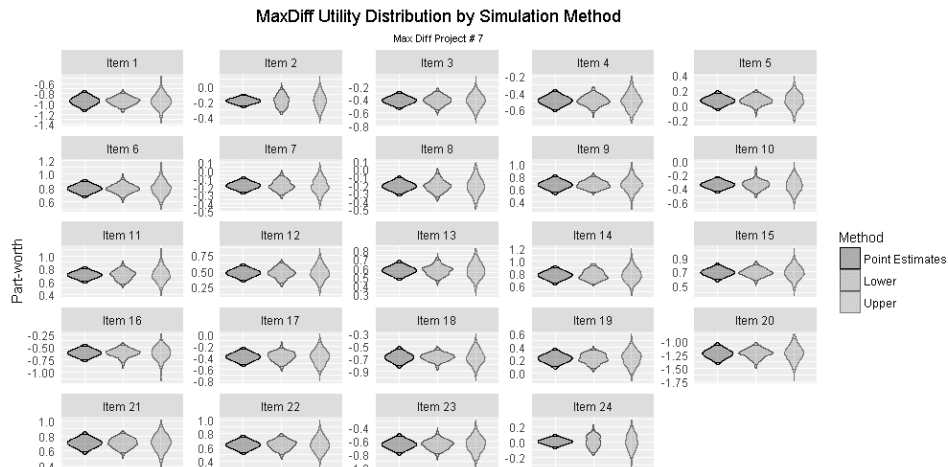


Each point in the figure represents a new share of preference simulated from the conjoint projects in the study. The x axis shows the point-estimate (frequentist) method for deriving share of preference, and the y axis represents the difference between the point-estimate share and the mean share from the posterior distributions. The figure demonstrates an interesting linear trend between the lower posterior distributions' mean share of preference and the point-estimate, demonstrating that the lower posterior distribution of shares tends to be more flat (closer to the mean share) than the point estimate shares.

MaxDiff Results—Utilities

Surprisingly, MaxDiff projects yielded very different results compared to conjoint, and much less alarming in terms of impact to inference. Figure 9 shows us the part-worth distributions on an example MaxDiff project. The figure demonstrates that *unlike* CBC projects, the distributions of part-worth utilities between estimation methods are very similar. However, like CBC projects, the point estimate method still tends to underestimate the variance, but to a much smaller degree.

Figure 9

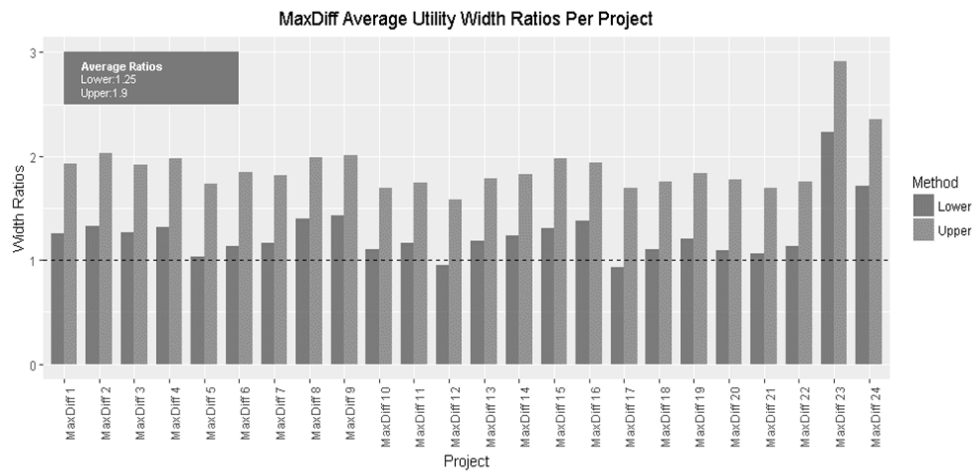


The figure shows part-worth distributions from frequentist (point-estimate) and posterior (upper, lower) distributions from an example HB model derived from a MaxDiff exercise. Each grid represents a MaxDiff item. We can observe that the frequentist distributions of part-worths tend to (mostly) underestimate the uncertainty found in posterior distributions. Compared to conjoint projects, however, the differences between distributions are more tame.

Figure 10 lays out the mean distribution width ratios for all MaxDiff projects. As seen in Figure 9, the point estimate distribution width is underestimated. On average, the point estimate

underestimated the lower level posterior distribution method by a factor of 1.25, and underestimated the upper level posterior distribution method of the upper method by 1.9.

Figure 10

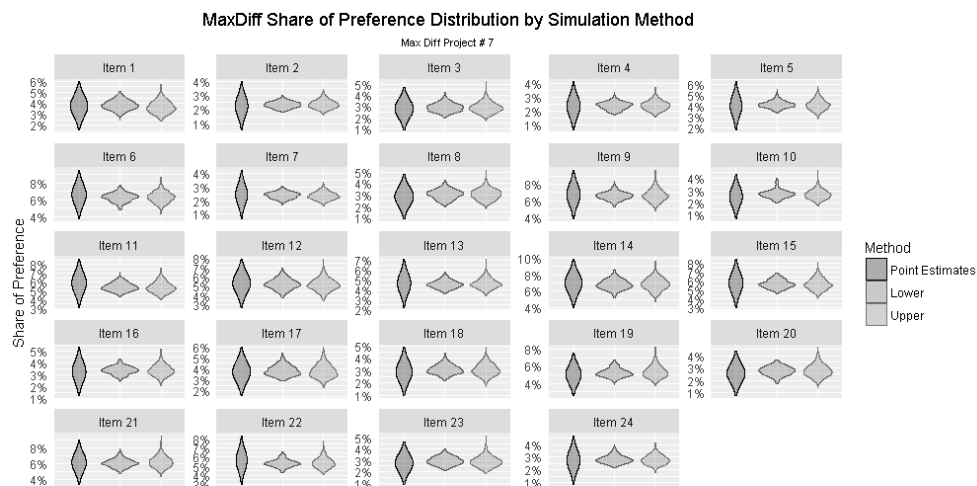


This figure represents the average "width" ratios from frequentist and posterior distributions of part-worths from multiple MaxDiff projects. For every model parameter in every project, we took 95% and 5% quantiles and measured the distances between them, in frequentist and posterior distributions. A ratio of 1 (represented by the dotted line) signifies the point where the width (uncertainty) of the frequentist and posterior (upper, lower) distributions are the same. We can observe the frequentist distribution width tends to be wider than posterior distribution widths. Compared to conjoint projects, however, the differences are smaller.

MaxDiff Results—Share of Preference

MaxDiff analysis measuring share of preference yielded something very different from CBC projects. Figure 11 shows an example mean share of preference distributions among the three estimation methods tested. One can observe here that all three share distributions for each MaxDiff item are approximately centered in the same place, and that the point estimate method overestimates the variance quite a bit.

Figure 11

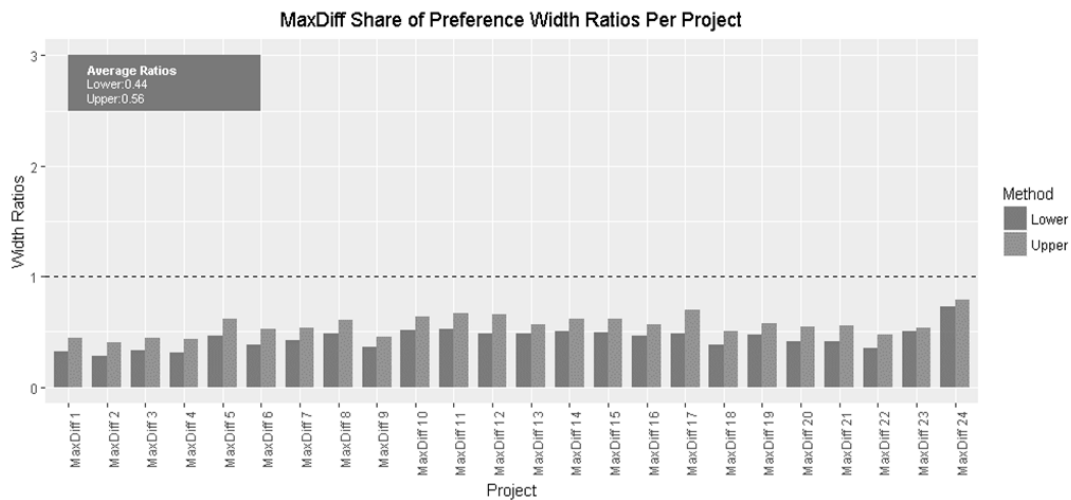


This figure shows share of preference distributions from frequentist and posterior (upper, lower) distributions from an example HB model derived from a choice based conjoint exercise. Each grid represents a MaxDiff item. We can observe that the share of preference distributions are centered in the same place, and that (unlike conjoint projects) the frequentist distribution tends to overestimate uncertainty compared to posterior distributions.

The overestimation of the point estimate method of MaxDiff items here can be observed in all of the MaxDiff projects in the study, which is demonstrated by Figure 12. The lower-level

posterior distribution method had an average distribution width ratio of 0.44, and the upper-level posterior distribution method, 0.56, when compared to the point estimate distribution.

Figure 12



This figure represents the average "width" ratios from frequentist and posterior distributions of share of preference from multiple MaxDiff projects. For every model parameter in every project, we took 95% and 5% quantiles and measured the distances between them, in frequentist and posterior distributions. A ratio of 1 (represented by the dotted line) signifies the point where the width (uncertainty) of the frequentist and posterior (upper, lower) distributions are the same. We can observe that in MaxDiff projects, the frequentist distribution tends to overestimate the uncertainty compared to posterior distributions

DISCUSSION

Our analysis shows that the shortcut point-estimate method for approximating the posterior distributions of utilities and shares of preference is substantially different than the upper-level or lower-level posterior distributions. Our data suggests that the point estimate method underestimates the variance around part-worth estimates, and often produces significantly different shares of preference in conjoint simulations. The shortcut method may be safer to use for MaxDiff simulations, as utility distributions are very similar, and share of preference distributions are conservative (wider than the lower or upper level posterior distributions). Conjoint studies tend to not only have very different distributions between methods, but mean shares of preference estimates are often significantly and substantially different from one another.

There are several ways this research could be extended and/or improved upon in the future. While we know that these methods produce different distributions of part-worth and share of preference. We did not cross-validate with true choice data, holdout tasks, or other prediction metrics to see which one was more accurate. Likewise, we did not compare other popular simulation techniques, like randomized first choice, which could put variance back into the part-worth estimates to make it more similar to upper and lower level posterior distributions. Adding meaningful covariates, especially in the upper level posterior distribution simulations, could have improved the quality of our models as well.

The point-estimate method could still be advantageous to the researcher, when the need for convenience, access to individual-level estimates, and the need to build simulators in Excel outweigh the need to estimate distributions of utilities and share of preference in a more theoretically consistent way. Researchers using any method should be careful when drawing

inference from these methods, and keep in mind that these methods may affect their research outcomes. When deciding between the upper- and lower-level posterior distribution when simulating preference share, it would be advantageous to consider the following questions:

1. Do you believe the sample to be representative of the overall population?
2. Are you trying to make inferences about these specific individuals or the population?
3. Did you include all the important upper level covariates to capture enough heterogeneity?

Due to the strong differences between simulation methods and their impact on inference, we believe that researchers should be making informed decisions about which method to simulate from. There are not only theoretical considerations, but practical considerations at play. Can your simulation tool accept the number of rows needed when using draws from the lower-level posterior distribution? Can your simulation tool filter to appropriate subsets of data in a reasonable amount of time? We encourage a holistic and informed approach when deciding on which method to use for generating insights fit for the business purpose.



Jacob Nelson



Edward Paul Johnson

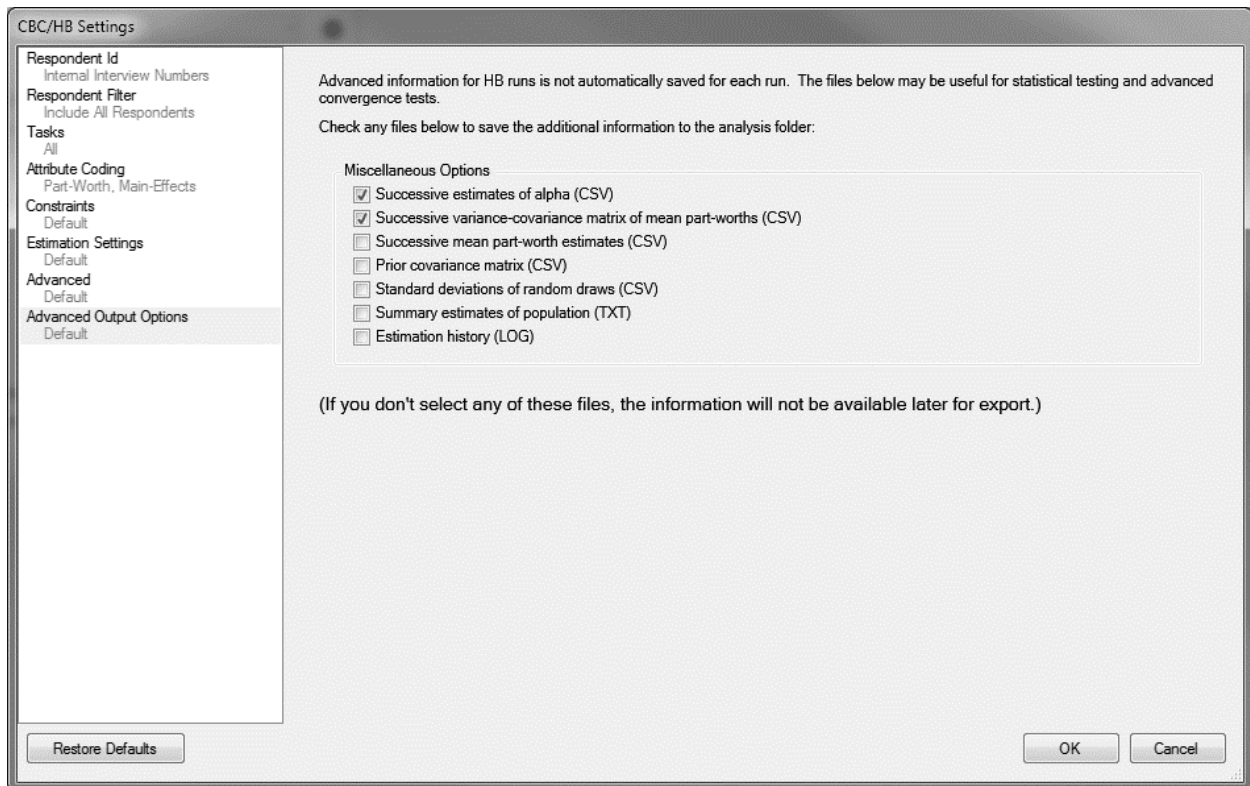


Brent Fuller

APPENDIX A—

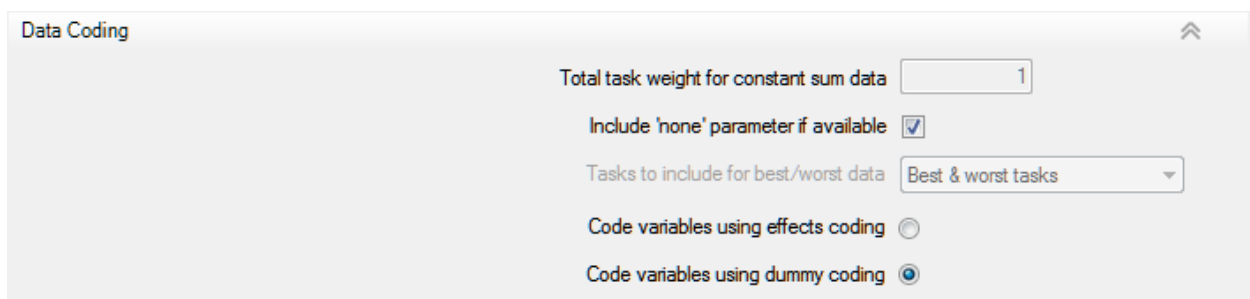
HOW TO SIMULATE RESPONDENTS FROM THE SAWTOOTH ALPHA FILE OUTPUT USING R

Before you estimate your HB model, make sure that you tell Lighthouse Studio that you want additional output files from the Advanced Output Options menu: “Successive estimates of alpha (CSV)” and “Successive variance-covariance matrix of mean part-worths (CSV).”



Note that if you are simulating for a MaxDiff project, these output files are not currently available in Lighthouse Studio (version 9.5.3). You will have to estimate in Sawtooth Software’s CBC/HB instead. You can do this by exporting a .CHO file, and importing it into CBC/HB.

For convenience, you should also tell your model to code variables using dummy coding, which you can do in the estimation settings. If you do not do this, you must dummy code the variables yourself.



Once your model is complete and you are satisfied with it, you'll be ready to simulate respondents in R. In this paper, we use some tidyverse-friendly approaches, but it should not be difficult to adapt to base-R techniques, if you prefer.

You will need to make sure that these packages are installed and loaded (all of which are available on the CRAN repository).

```
library(readr)
library(dplyr)
library(purrr)
library(tidyr)
library(tibble)
library(metaSEM)
library(mvnfast)
```

Once the packages are loaded, you will need to load your alpha and covariance csv files into R.

```
alpha_raw <- read_csv("<PATH TO YOUR ALPHA CSV FILE>")
covariance_raw <- read_csv("<PATH TO YOUR VARIANCE-COVARIANCE.CSV FILE>")
```

These files will need to be cleaned before we can simulate from them. Each row of both the alpha and covariance data frames represent a draw, and contains every iteration used in the model, including the burn-in iterations (but also skipping per the skip factor that was set in the estimation settings). We will need to filter the data to include only the converged draws. You may decide to only use some of the converged draws to make this problem easier on your computer. For this example, we will use 5000 converged draws. We will also get rid of the "Iteration" column from both data frames, and all the holdout variables from the alpha data (variables where the dummy-coded part-worths are equal to 0), as they are not used.

```
keep_converged <- 5000

alpha <- alpha_raw %>%
  tail(keep_converged) %>%
  select(-Iteration, -which(map_lgl(., ~all(.x == 0))))

covariances <- covariances_raw %>%
  tail(keep_converged) %>%
  select(-Iteration)
```

Each row of the covariance file that Sawtooth Software produces represents a flattened symmetric matrix for each draw, so they need to be unflattened. This is done by looping through each row of the covariances data frame and applying the "vec2symMat()" function from the metaSEM package to each row (Cheung, 2015). Each matrix will be nested in the data frame.

```
sym_covariances <- pmap(covariances, ~vec2symMat(c(...))) %>%
  tibble(sigma = .)
```

We can now combine the alphas and symmetric covariance matrices into a single data frame with more nesting, which we will name "upper_draws."

```
upper_draws <- alpha %>%
  group_by(draw = 1:n()) %>%
  nest(.key = mu) %>%
  ungroup() %>%
```

```
bind_cols(sym_covariances)
```

If everything has executed correctly, you should have a two-variable data frame, containing the nested part-worths (μ) and the nested covariance matrices (σ). Each part-worth observation should have the same number of columns as the covariance matrix in observation. Printing the upper draws object to the console will look like this (depending on your data).

```
upper_draws
# A tibble: 5,000 x 3
draw mu sigma
<int> <list> <list>
1 1 <tibble [1 x 22]> <dbl [22 x 22]>
2 2 <tibble [1 x 22]> <dbl [22 x 22]>
3 3 <tibble [1 x 22]> <dbl [22 x 22]>
4 4 <tibble [1 x 22]> <dbl [22 x 22]>
5 5 <tibble [1 x 22]> <dbl [22 x 22]>
6 6 <tibble [1 x 22]> <dbl [22 x 22]>
7 7 <tibble [1 x 22]> <dbl [22 x 22]>
8 8 <tibble [1 x 22]> <dbl [22 x 22]>
9 9 <tibble [1 x 22]> <dbl [22 x 22]>
10 10 <tibble [1 x 22]> <dbl [22 x 22]>
# ... with 4,990 more rows
```

At this point, you are ready to begin simulating respondents. You will need to decide how many simulated respondents you will want to simulate for each draw. For this example, we will simulate 1000 respondents. Depending on the complexity of the data, you may need more simulated respondents if you want each draw to converge to the mean.

```
n_resp <- 1000
```

Simulating respondents is done by drawing from the multivariate normal distribution via the `rmvn()` function from the `mvnfast` package (Matteo, 2016), given μ and σ for each draw.

```
simulated_respondents <- pmap_dfr(upper_draws, function(draw, mu, sigma) {
  rmvn(n = n_resp, mu = as.numeric(mu), sigma = sigma) %>%
  as_tibble %>%
  add_column(draw = draw, sim_resp = 1:n_resp, .before = 1)
})
```

This may take several minutes to run. Note that this operation may require significant amounts of memory (this example, for instance, will result in a data frame of 5 million rows). If your computer resources are limited, you may need to split your upper draws object into batches, or you may need to reduce the number of draws or simulated respondents you use.

After you are done simulating respondents, your final data frame should contain a draw index variable, a simulated respondent index variable, and simulated part-worth scores (dummy coded) from the multivariate normal distribution.

WORKS CITED

- Allenby, Greg M. and James L. Ginter (1995). Using Extremes to Design Products and Segment Markets. *Journal of Marketing Research*, 32, 392–403.
- Allenby, G.M., Brazell, J.D., Howell, J.R. et al. (2014) Economic valuation of product features. *Quant Mark Econ* (2014) 12: 421. <https://doi.org/10.1007/s11129-014-9150-x>
- Allenby, G.M., & Rossi, P.E. (2003). Perspectives based on 10 years of HB in marketing research. In *Sawtooth Software Conference Proceedings* (pp. 157–169).
- Chapman, C., & Feit, E.M. (2015). *R for marketing research and analytics*. Cham: Springer.
- Cheung, M.W.L. (2015). {metaSEM}: An R Package for Meta-Analysis using Structural Equation Modeling. *Frontiers in Psychology* 5, 1521. URL <http://journal.frontiersin.org/article/10.3389/fpsyg.2014.01521/full> DOI 10.3389/fpsyg.2014.01521
- Kurz, P., & Binner, S. (2016). Simulating from the HB upper level model. *2016 Sawtooth Conference Proceedings*.
- Lee, Jake, (2016). Conjoint Upper-level HB Models: Benefits and Practical Tips. *2016 ART Forum*.
- Matteo Fasiolo. An introduction to mvnfast (2016). R package version 0.1.6.
- Morris, C.N. (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381), 47–55.
- Sawtooth Software Inc. Introduction to Market Simulators for Conjoint Analysis (2009). <https://www.sawtoothsoftware.com/support/technical-papers/market-simulations/introduction-to-market-simulators-for-conjoint-analysis-2009>.
- Sawtooth Software (2009). The CBC/HB System for Hierarchical Bayes Estimation Version 5.0 Technical Paper. <https://www.sawtoothsoftware.com/support/technical-papers/hierarchical-bayes-estimation/cbc-hb-technical-paper-2009>

PREFERENCE-BASED CONJOINT—CAN IT BE USED TO MODEL MARKETS WITH MANY DOZENS OF PRODUCTS?

JEROEN HARDON
MARCO HOOGERBRUGGE
SKIM

SUMMARY

In this paper we explore two alternative conjoint approaches that aim to predict better in a situation when we have dozens of products in the simulator—but not so many on the screen during the interview. Both approaches are a significant improvement over current practices. However, more work in this area is still needed.

INTRODUCTION

In a number of markets, for example in the services industries and the technology industries, every operator or manufacturer has a large assortment of different products that potential customers can choose from. For research based on Choice-Based Conjoint this often leads to market simulators in which many dozens of products are listed. However, this conflicts with the design of the conjoint exercise in which (merely due to screen capacity) we typically show three or four products. The result of this discrepancy between interview environment and analysis environment is that our predictions are certainly inaccurate and perhaps also biased. As we will see later, with a few dozens of products in the simulator, we are only able to predict the correct preferred product (based on first choice) for some 20%–30% of the respondents. In this situation there must be a lot of improvement possible.

Intuitively, Adaptive Choice-Conjoint (ACBC) should be able to better cope with the situation that we described. The occurrence of levels in ACBC differs per individual respondent, dependent on a prior Build-Your-Own (BYO) task. The result is that the respondents much more often evaluate attribute levels that are relevant to them personally during the choice tasks. For example, if a respondent really wants to have something from brand A (as indicated in the BYO task), then brand A is being shown more in the conjoint exercise and we will better know which product the respondent will take *within* the assortment of brand A. However, as we will see later, in practice it does not always work better with dozens of products in the simulator. Our hypothesis about ACBC is that it is too extreme in its execution and it does not give a good enough insight in realistic trade-offs that respondents have to make. This can, for example, be due to the fact that respondents actually have preference for 2 levels of an attribute (e.g., 2 brands), which is not captured in the BYO task.

The two new ideas that we will discuss in this paper are:

1. Preference-Based Conjoint (PBC), which is conceptually a variation of ACBC approach where we have unlimited flexibility to tune the occurrence of attribute levels. We tested a variant in which we considerably decreased the occurrence of the most preferred level (in comparison in ACBC), considerably increased the occurrence of the adjacent levels of the

most preferred level and kept the low occurrence of the levels that are far from the respondent's preferred level.

2. PBC² (pronounce PBC square), which is a variation of Choice-Based Conjoint—as it does not make use of a prior BYO task—it increases the probability of the attribute levels being shown based on the actual responses during the choice tasks. In other words, it is even more an on-the-fly approach. We named it “square” because during the interview it *increasingly* adjusts based on the responses: in the last task the respondent's preferred levels are much often more shown than in the beginning, because in the beginning we do not know a lot about these preferences.

PREFERENCE-BASED CONJOINT

While conceptually PBC is more like a variation of ACBC, technically it is executed in a CBC environment. Suppose we have six levels of data allowance (in a mobile subscription study), we *program* for example eight levels, with 6 *fixed* levels, varying from no data to unlimited data, and make a reservation for two *flexible* levels of which the text is any of the 6 levels, dependent on the respondent's initial preference (as expressed e.g., in a BYO task). In the end, we are using a complete enumeration design with 8 levels for this attribute:

- 6 fixed levels
- 2 flexible levels for the respondent's preferred level

Suppose level 3 was the preferred level, then the chance of level 3 being shown in the choice tasks is three times higher, because the text of level 3 shows up when the complete enumeration design indicates level 3, 7 or 8.

Note that we are entirely flexible in our survey design. We might alternatively define 15 levels:

- 6 fixed levels
- 3 flexible levels for the respondent's preferred level
- 2 flexible levels for the respondent's preferred level +1
- 2 flexible levels for the respondent's preferred level -1
- 1 flexible levels for the respondent's preferred level +2
- 1 flexible levels for the respondent's preferred level -2

Note, in addition, that the determination of the preferred level does not *have to* be based on a BYO task (in ACBC it has to). Alternatively, we can also assume preferred levels from various separate select questions in the survey. However, in the test studies that we did we used a BYO task after all, just in order to have a fair comparison with ACBC.

An additional (assumed) advantage of PBC over ACBC is that we are not limited to a certain amount of concepts per task. ACBC only allows 2 or 3 concepts per task.

PBC²

The idea behind PBC² is the same as for Preference-Based Conjoint but the “feeding” of the flexible levels is entirely different. So again we may have:

- 6 fixed levels
- 3 flexible levels varying per respondent

But in this case the flexible levels are not dependent on any prior questions before the conjoint exercise. They are instead dependent on the choices *in* the conjoint exercise. So it becomes, for example in task 4:

- 6 fixed levels
- Flexible level, namely the level of the chosen concept of task 1
- Flexible level, namely the level of the chosen concept of task 2
- Flexible level, namely the level of the chosen concept of task 3

The amount of flexible levels increases in the course of the interview, for example in task 10 we have 6 fixed levels and 9 flexible levels for 15 levels in total. So in task 10, in the most extreme case (which is highly unlikely), one level will have a probability to occur of 10/15 and the other five levels 1/15 each.

Just for ease of processing, we implemented PBC² as follows, in four series:

- No flexible levels in task 1–3
- 3 flexible levels in task 4–6 (the levels of the chosen concepts of task 1–3 are duplicated)
- 6 flexible levels in task 7–9 (the levels of the chosen concepts of task 4–6 are duplicated)
- 9 flexible levels in task 10–12 (the levels of the chosen concepts of task 7–9 are duplicated)

Note that PBC² has a strong theoretical advantage over ACBC *and* PBC, in the sense that it does not require making any assumptions/choices in the design. In ACBC and PBC one has to determine upfront *for which attributes* the frequencies of the levels depend on earlier answers. Consequently we may vary the frequencies of an attribute that appears to be entirely unimportant, or we may not vary the frequencies of an attribute that is important after all. In PBC², if an attribute turns out to be unimportant, its frequencies will remain evenly distributed, simply by the nature of the algorithm.

FIRST TEST STUDY

Introduction

We have conducted two test studies in the same market, of mobile telephony subscriptions in the Netherlands, in early 2017 and early 2018. In this period the prices of the subscriptions have decreased substantially.

We used 7 attributes with the same levels in both studies, but in the second study we adjusted the component prices to be in line with the changing market.

- Brand (6 levels)
- Data allowance (7 minutes)
- Minutes allowance (6 levels)
- Data out-of-contract usage (3 levels)
- Contract period (3 levels)
- Expiration of allowance (3 levels)
- Summed pricing¹ +/-30%

The Holdout Task

The biggest challenge was how to define a *holdout task*. After all, the aim is to predict a simulator with dozens of products well. So the holdout task should somehow represent a simulator that contains dozens of products. In the first test study, we defined a holdout task with 20 products, omitted a number of attributes (so it was a partial profile holdout task) and showed the products on one screen in two columns, sorted by brand, in order by make. In retrospect we tend to believe that the sorting by brand may well have been a bias factor, making brand more/too important. Therefore we will discuss the first test study more briefly than the second test study.

Figure 1. Holdout Task 1st Test Study

Which product would you choose?

• T-Mobile, 300 minuten, 1 GB, €15	• Vodafone, 150 minuten, 5 GB, €24
• T-Mobile, 120 minuten, 6 GB, €22	• Vodafone, onbeperkt bellen in NL, 1,5 GB, €26
• T-Mobile, onbeperkt bellen in NL, 3 GB, €24	• Tele2, 100 minuten, 1 GB, €10
• T-Mobile, onbeperkt bellen in NL, 6 GB, €27	• Tele2, onbeperkt bellen in NL, 4 GB, €14
• KPN, 200 minuten, 1 GB, €20	• Tele2, onbeperkt bellen in NL, 8 GB, €17
• KPN, onbeperkt bellen in NL, 1 GB, €25	• Tele2, onbeperkt bellen in NL, 24 GB, €26
• KPN, onbeperkt bellen in NL, 5 GB, €30	• Ben, 100 minuten, geen databundel, €7
• KPN, onbeperkt bellen in NL, 10 GB, €40	• Ben, 100 minuten, 1 GB, €10
• Vodafone, 150 minuten, geen databundel, €7,50	• Ben, 300 minuten, 1 GB, €11
• Vodafone, 150 minuten, 1 GB, €15	• Ben, onbeperkt bellen in NL, 1 GB, €16

The Test Legs

We had test legs of 250 respondents per leg. In all test legs we had a Build-Your-Own task, even if we didn't use that task for the generation of the design (like in CBC and PBC²). Furthermore, while summed pricing is the default in ACBC, we have implemented the summed pricing approach in CBC, PBC and PBC² as well, on the one hand for fair comparison, and also because it is entirely realistic in the market. The ACBC leg as described below did not contain a screening exercise, which implied that all respondents evaluated a full set of choice tasks.

¹ For more information about summed pricing, see for example <https://www.sawtoothsoftware.com/help/lighthouse-studio/manual/priceinadaptivecbc.html>

In ACBC, PBC and the default PBC² leg we only varied the frequencies of three attributes: of data, minutes, and expiration of allowance, and we only had these attributes in the BYO task including the summed price total. Note that we did this for PBC² *only* for comparison reasons, because as noted earlier the advantage of PBC² is that we do not need to establish upfront which attributes have levels with varying frequency. In addition we also had another leg with PBC² in which levels of all attributes were allowed to vary in frequency.

The Results

We evaluated the study based on hit rate and mean absolute error (MAE) in the holdout task, in three different ways:

- Utilities based on HB only based on choice tasks, without covariates
- Utilities based on HB based on choice tasks and BYO task, without covariates
- Utilities based on HB based on choice tasks and BYO task, with current brand used as a covariate

In all cases we run HB with a prior variance of 0.5 and 5 degrees of freedom and we used the point estimates. Furthermore we run HB in three replications with different starting seeds each time, and took the average hit rate and average MAE across the replications. We have done this after we noted that every replication resulted in a different hit rate and MAE, even with huge amounts of iterations.

The outcomes are as follows:

Table 2. Hit Rates (Average Across 3 Replications)

	HB choice tasks only	HB choice tasks + BYO	HB choice tasks + BYO, with covariate
CBC	19.6%	23.6%	29.4%
ACBC (without screening)	19.0%	19.8%	27.3%
PBC	24.8%	26.2%	31.9%
PBC ² (varying level freqs of same 3 atts as ACBC and PBC)	21.9%	22.5%	28.4%
PBC ² (varying level frequencies of all attributes)	28.0%	28.5%	32.6%

Table 3. Mean Absolute Error (Average Across 3 Replications)

	HB choice tasks only	HB choice tasks + BYO	HB choice tasks + BYO, with covariate
CBC	1.8%	1.7%	1.6%
ACBC (without screening)	1.8%	1.8%	1.8%
PBC	1.3%	1.2%	1.2%
PBC ² (varying level freqs of same 3 atts as ACBC and PBC)	1.9%	1.9%	1.8%
PBC ² (varying level frequencies of all attributes)	1.2%	1.2%	1.1%

By and large a higher hit rate correlates with a lower MAE, which is very fortunate for drawing conclusions about the different methods. In all cases the second PBC² leg is the “winner” but *only* when allowing *all* attributes to vary in the design. PBC (allowing only 3 attributes to vary in the design) follows closely. This led us to believe that we could improve PBC by varying more attributes in the design. Note that good old CBC did not perform badly at all when looking at hit rate, but relatively worse when looking at MAE.

The figures are not only interesting for the sake of comparing the different legs. We can also draw some conclusions across the legs. First of all, adding BYO data to the utility estimation (even if the BYO data hadn’t been used in the design, in CBC and PBC²) adds a little value in the predictions but not so much. On the other hand, adding the current brand covariate has a huge impact and this is an effect that we have not seen before. Here we were getting a bit suspicious about the set-up of our holdout task: could we artificially have increased brand loyalty in the holdout task by sorting the concepts by brand?

The figures as in the tables are not the whole picture though, because when we analyzed the MAE calculations in more detail we saw that *all* research legs were biased in the sense that the actual *counts* of the holdout tasks revealed that respondents chose more expensive products than had been *predicted*. This is a finding that returns in our second test study. We also found significant deviations between counts and predictions when we looked at aggregate shares at brand level. So deviations between actual choice in a 20-concept holdout task and predictions based on 3 or 4 concepts per task is not only a matter of accuracy but also of a matter of bias.

SECOND TEST STUDY

The Holdout Task

As noted, in hindsight we were a bit hesitant about the holdout task in the first test study. For the second test study we developed an alternative holdout procedure. First of all, we now had 40 concepts in the holdout instead of 20 which is more realistic in the sense that actual market simulators also often have 40 or more concepts. Second, we split the holdout exercise into three questions:

1. For 20 concepts we asked which concepts the respondent would consider buying (multiple choice).
2. On a next screen, for 20 other concepts, we asked the same question.
3. On the last screen, we only showed the concepts that the respondent would consider buying and we asked which one they would most likely buy (single choice).

For sorting of the concepts in this exercise, we split each leg randomly in half:

- For half of the respondents we first sorted on brand: 3 brands with 22 concepts on the first page, 3 brands with 18 concepts on the second page.
- For half of the respondents we first sorted on “tier”: 20 low tier products sorted by brand on the first page, 20 high tier products sorted by brand on the second page
- On the final page with only the evoked set of the respondent, we randomized the order. Typically respondents chose 2 to 7 products in their evoked set. So despite randomization this was still a doable task for respondents, and because of the randomization in this final task we hoped to avoid bias toward brand.

The Test Legs

In this case we also had an ACBC leg with screening exercise (including two unacceptable questions), just for comparison. Most respondents in this leg had a reduced number of choice tasks, in case they rejected concepts or attribute levels.

For PBC², we no longer included the variant with only three attributes to be flexible in level frequency, since in the first study the variant with all attributes to be flexible in level frequency outperformed the former variant so much. So all attributes were allowed to vary in level frequency.

For PBC, we also included brand as an attribute to be flexible in level frequency. We did that because we saw that PBC² performed so much better when having flexible level frequency with all attributes included. We presumed (probably wrongly, as we will see) that brand would be the most important attribute to be added. So on behalf of PBC we included the brand attribute in the BYO exercise, and we did that also in all other legs.

How Did the Level Frequencies Work Out in This Test Study?

In PBC and even more in ACBC we have enforced that one brand appears more often in the choice tasks. However, when we checked the PBC² data we noticed that brand level frequencies were pretty evenly distributed for a large amount of respondents. So in hindsight it may not have been a good idea to add brand to the BYO task and enforce the most preferred brand level to occur more often. (The question then becomes how PBC² can reasonably predict the right brand in the holdout task, the answer must be that this is due to adding the BYO data as conjoint tasks in the utility estimation.)

For minutes and data we obviously see that especially in ACBC a certain level of minutes and data is shown more often. The distribution in PBC is more flat (as intended), except when a respondent chooses the very lowest or very highest level of these attributes. In PBC² we see an interesting phenomenon: the frequency of the adjacent higher level of data and minute is a lot

more than the adjacent lower level of data and minutes (we mean adjacent to the BYO level). Apparently respondents are more prepared to give in on price than giving in on minutes and data, comparing to the BYO task. With ACBC and PBC we cannot foresee effects like this, while PBC² adjusts naturally to this phenomenon.

We looked at the occurrences of price, and for that purpose we set apart two sets of respondents who answered one particular combination of brand, minutes and data in the BYO task. One set consists of certain “mid-high end” respondents and the other of certain “low end” respondents. In Table 4 we showed for each of the methods the “typical” price range of the concepts shown to these respondents. In this range there is a close to uniform distribution while outside this range the relative frequencies reduce quickly. We have established this “typical” range by face value, looking at the histograms, so these values are not very precise. Nevertheless two important phenomena immediately become clear:

1. All adaptive methods (ACBC, PBC and PBC²) contain a huge price range for the middle-high end respondents, hardly any better than CBC, while the latter is non-adaptive.
2. ACBC tests more concepts that are below the BYO price for the low end respondents and tests more concepts that are above the BYO price for the high end respondents. But the flipside of this is that ACBC has an even wider price range.

Table 4. “Typical” Price Range in Choice Tasks (For Two Types of Respondents)

Method	Example “mid-high end” respondent: in BYO brand #2, 10 GB, unlimited minutes, €29	Example “low end” respondent: in BYO brand #5, 1 GB, 100 minutes, €8
CBC	€5–€25	€5–€25
ACBC (without screening)	€8–€40 (but with a pretty steep valley in the range of €22–€26)	€4–€18
PBC	€10–€32	€8–€14
PBC ²	€8–€28	€8–€18

In any case, this suggests that there must be huge room for improvement for PBC and especially for PBC² beyond what we discuss in this paper, such that they keep the prices in the choice tasks 1) in a *narrower* range around the BYO price and 2) in a range that is *equally under and above* the BYO price level. For ACBC the former point of improvement is even more important, while the latter point of improvement has already been reached a bit. By and large we may conclude that currently none of the methods is very adaptive in terms of the prices being shown in the choice tasks.

The Results

The outcomes are as follows:

Table 5. Hit Rates (Average Across 3 Replications)

	HB choice tasks only	HB choice tasks + BYO	HB choice tasks + BYO, with covariate
CBC	18.0%	24.6%	26.9%
ACBC (with screening and unacceptables)	20.4%	28.1%	27.2%
ACBC (without screening)	23.5%	31.4%	29.8%
PBC	22.9%	26.4%	27.2%
PBC ²	23.2%	31.3%	32.0%

Table 6. Mean Absolute Error (Average Across 3 Replications)

	HB choice tasks only	HB choice tasks + BYO	HB choice tasks + BYO, with covariate
CBC	1.2%	1.1%	1.1%
ACBC (with screening and unacceptables)	1.1%	1.1%	1.1%
ACBC (without screening)	1.1%	0.9%	0.9%
PBC	1.1%	1.0%	1.0%
PBC ²	1.0%	1.0%	1.0%

Like in the first test study, PBC² is the “winner.” The improvement that we hoped to reach for PBC (by including brand as an attribute with flexible level occurrences) does not occur at all, rather on the contrary. Also ACBC without screening is performing quite well, while ACBC with screening (but mostly fewer choice tasks) performs clearly worse than ACBC without screening. We have no explanation why ACBC without screening performs so much better in the second test study than in the first test study (except for sampling error, to be discussed at the end of this presentation).

We can also draw some conclusions across the legs. First of all, adding BYO data to the utility estimation (even if the BYO data hadn’t been used in the design, in CBC and PBC²) adds much more value in the predictions than in the first study. We believe this is due to the difference in brand signal. In the first study brand was not a part of the BYO, and we do not see a large improvement when adding the BYO to the data. But we do when adding brand as a covariate. In the second study (where brand was part of the BYO) we see the impact occurring when adding the BYO data already, concluding that adding more brand signal either via the BYO or the covariates helps the prediction of brand choice in the holdout task.

The figures as in the tables are not the whole picture though, because when we analyzed the MAE calculations in more detail we saw that *the CBC and ACBC* research legs were biased in the sense that the actual *predicted choice* of the holdout tasks revealed lower prices than had been *actually chosen*. The results of this are shown in the table below.

Table 7. Percentage of Respondents for Whom the Price of the Predicted Choice in the Holdout Task Was Higher/Lower Than the Actual Choice in the Holdout

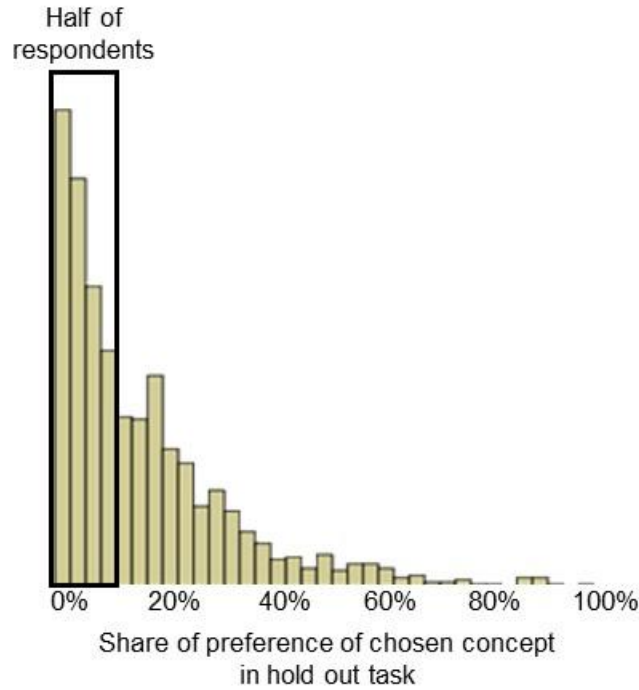
	CBC	ACBC with screening	ACBC without screening	PBC	PBC ²
Underestimated	39.1%	36.4%	31.0%	35.9%	33.2%
On par	29.8%	31.6%	36.7%	29.4%	34.0%
Overestimated	31.0%	32.0%	32.2%	34.7%	32.8%
Average price difference (in €)	-0.30	-0.60	-0.40	0	+0.10

By the way, the table above does not mean to suggest that the predictions of PBC and PBC² were *overall* unbiased. We just happen to look here at the results from one particular angle, namely price. There were biases for PBC and PBC² as well, when we analyzed aggregate brand shares or by data allowance, but they were not pointing in any particular consistent direction. This phenomenon may well have a mathematical rather than a psychological cause (see paper “A Meta-Analysis on Three Distinct Methods Used in Measuring Variability of Utilities and Preference Shares within the Hierarchical Bayesian Model,” by Jacob Nelson, Edward Paul Johnson, and Brent Fuller, in these same proceedings). So the above results are nice, but not conclusive.

Looking At an Alternative Metric

A hit rate of 25%-30% sounds like a very good performance, taking into consideration that the holdout task consists of 40 products. Random data would result in a hit rate of only 2.5%. However, when we look at it in more detail, by checking the *share of preference* of the chosen concept in the holdout task, it appears that with no less than half of the respondents that share is below 10% which implies that it was a pretty bad prediction. With 10% of the respondents the share of preference is even below 1% which implies an awfully bad prediction.

Figure 8. Histogram of Shares of Preference of the Chosen Concept in the Holdout Task of All Legs Combined



A better, and also theoretically sounder, metric to capture these differences in quality of prediction, is by looking at the geometric mean of these shares of preference. The percentages then become a lot lower, around 8%, which is also closer to the median share of preference. The conclusions about the relative performance of PBC² versus ACBC versus PBC versus CBC did not change when using this measure, but the relative magnitude of the difference increases. In particular PBC² outperforms better than before, i.e., is better able to prevent exceptionally bad predictions. And anyway we think it is important to share this little piece of analysis in order to emphasize that there is a BIG gap between 8% and 100% that we should still improve on, somehow.

Table 9. Traditional Hit Rate Versus Geometric Mean of Share Predictions of the Chosen Concept

Leg	Hit rate (with covariates)	Geometric mean SoP (with covariates)
CBC	26.9%	7.1%
ACBC with screening	27.2%	7.3%
ACBC without screening	29.8%	8.1%
PBC	27.2%	7.6%
PBC ²	32.0%	9.2%

COMPARING THE RESULTS OF THE TWO STUDIES

Across the two studies, PBC² performs best and plain CBC does not perform as well. Those conclusions can be drawn without much doubt. With the other methods there is more nuance needed.

PBC has been tested in two slightly different variants in the two studies and we may well conclude that the set-up of the second study, with *additionally* including brand as a flexible attribute, was not as successful. On the other hand PBC² has been tested in two variants within the first study and inclusion of *all* attributes as flexible attributes was more successful there. This sounds quite contradictory. The problem is that PBC *enforces* a different probability distribution of level occurrences *for every respondent*, even if brand is largely irrelevant for a particular respondent. On the other hand, PBC² only changes the probability distribution of level occurrences *if* it becomes clear during the survey that an attribute is a relevant choice criterion. So brand may an important criterion for a couple of respondents, or it may be one of the three last “less important” attributes for a couple of respondents, and PBC² will then take that into account on an individual respondent basis (while PBC has kept the last three attributes equally distributed for all respondents). So PBC² is much more flexible and much better individualized than PBC.

ACBC without screening performs relatively poorly in the first test study and relatively well in the second test study. We do not have a good explanation for this phenomenon. The only thing that we can say about it, for now, is that there is also some measurement error involved in our estimations, which may make the results vary anyway. So that could be the reason for the deviation between the two studies. Measurement error may be divided in two components:

1. *Sampling error*. Every leg contained 250 respondents which results in 95% confidence intervals for the hit rates of +/-2% (total width 4%). That is a pretty big range and so we may just have had bad luck in the sense that the point estimate of hit rate of the ACBC leg may have been at the lower end of the confidence interval in the first study and at the higher end of the confidence interval in the second study. In retrospect, 250 respondents is not an awful lot.
2. *HB estimation error*. We have noticed that HB does not globally converge when there are constraints in play (not even with 100,000 iterations). Our procedure to come around this problem was as follows: we have taken three replications for each leg (with different starting seeds), calculated hit rate and MAE of that leg, and in the end calculated the *average* hit rate and MAE of those three replications. From a practical perspective this doesn't feel too good, because we normally only do one replication. What should we now do in practice, take averages of multiple replications like we did here, or take the replication with the highest hit rate across multiple replications? More research is needed here.

Just as an illustration of this phenomenon, Table 10 shows 6 replications for two legs.

Table 10. Hit Rate in Different Replications (With Different Starting Seeds) with 20,000 Burn-in Iterations and 30,000 Saved Iterations

Starting seed	Hit rate ACBC without screening	Hit rate PBC
1	28.6%	26.1%
2	30.2%	27.3%
3	32.7%	26.9%
4	29.8%	24.9%
5	29.0%	24.9%
6	31.8%	25.7%

POSTSCRIPT 1

During the Sawtooth Software Conference we fielded another 250 respondents with an alternate version of PBC². Instead of *duplicating* the levels of concepts that were chosen by respondents, we even *triplicated* these levels. So this was a variant with a steeper rate of adaptation to the individual respondents' choices. We were aiming at addressing the problem that we described in the section "*How did the level frequencies work out in this test study?*"

Unfortunately, on the one hand the price range of the concepts moved only a little:

- For the mid-high end respondent the price range moved from €8–28 to €10–30.
- For the low end respondent the price range from €8–18 did not move at all.

Also, when testing the survey, we noticed that we might move into a respondents' evoked set more quickly. But on the other hand it also happened, accidentally, that levels of irrelevant attributes were chosen more often and then the number of occurrences of that level exploded quickly. This is due to the fact that our algorithm is based on absolute frequency of levels (picked) rather than relative frequency (picked/shown).

Finally, probably as a result of the latter, hit rate was somewhat lower than regular PBC². So this is a direction that we do not need to explore any further. The question remains how we can tweak the design such that more concepts are being shown to an individual respondent that are in his/her preferred price range?

POSTSCRIPT 2

After we delivered the presentation at the Sawtooth Software Conference, Bryan Orme came to us and said: “You had a similar idea (to PBC²) some 10 or 15 years ago, don’t you remember?” Honestly at that point of time we didn’t remember, but in the meantime some of our memories came back.

First of all, it was an idea that *failed* (based on hit rate), so we never presented it, and that is probably the reason we forgot.

Second, it had been implemented as a direct *replacement* of attribute levels, rather than by changing the *probability of occurrence* of attribute levels. So for example in the 5th task we replaced the levels of attributes 1 and 5 by the level in the winning concepts of the 1st to 4th task. And in the 6th task we replaced levels of attributes 2 and 4. And so on. In retrospect it is difficult to judge whether changing the probability of occurrence is better or worse than direct replacement.

Third, we most probably tried this before ACBC was introduced, because it had been implemented in Ci3, and so most probably we treated price the same as any other attribute and alternated in replacing price levels of winning concepts throughout the exercise. In retrospect that may have been a likely reason for failure, because on the one hand products were getting more attractive content for a respondent in the course of the survey, but *also* potentially were getting less expensive during the survey (or at least getting more value for money). This may well have converged to concepts that were “too good to be true” and would never exist in the real market. On the other hand in ACBC, PBC and PBC², when applying summed pricing, price is being varied *independently* of the consumer’s previous choices (while price does depend on the attribute levels) and that may work out much better.

Looking back, while it had never so explicitly been promoted, we think that (the easy way of) applying summed pricing has actually been one of the most important innovations of ACBC. In practice, the idea of summed pricing has drastically reduced the design space: in old-school CBC applications with price as an independent attribute we often just had some prohibitions with price, reducing the design space by for example 30%, while with summed pricing we reduce the design space by for example 80%, thus *limiting the design space more drastically to products that are more realistic in the actual market*. The latter is one key element of predicting the right choice in a simulator with dozens of products, even while much more work is needed in that area.

CONCLUSIONS

1. The issue of having dozens of products in the simulator (while having only a few concepts per screen in the survey) has not been explored in depth so far. We hope, by means of this paper, to have made a start in this area.
2. We are glad that one of our ideas to cope with the problem performed better in the simulator than the standard solutions that are available on the market.
3. Nevertheless we do not have the illusion that we have really solved the problem. A geometric mean SoP of 9.2% is an improvement compared to standard solutions, but still low.

4. One point we may definitely improve upon is that we should show concepts to a respondent in a *price range* that is more relevant for him/her. Despite the adaptiveness of the design algorithms (except in CBC), the prices of different concepts varied hugely for a single respondent, and besides the range was biased: for high end respondents we mainly show concepts that are less expensive than in the BYO, for low end respondents we mainly show concepts that are more expensive than in the BYO. ACBC performed somewhat better in preventing this bias, but only because the price range of the shown concepts was even wider.



Jeroen Hardon



Marco Hoogerbrugge

DEVELOPMENT OF AN ADAPTIVE TYPING TOOL FROM MAXDIFF RESPONSE DATA

JAY MAGIDSON

STATISTICAL INNOVATIONS

JOHN P. MADURA

UNIVERSITY OF CONNECTICUT AND

STATISTICAL INNOVATIONS

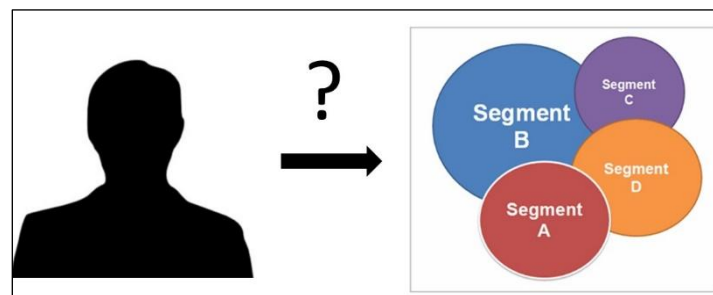
ABSTRACT

A new adaptive approach for developing MaxDiff typing tools achieves high accuracy with only 8 binary comparisons (tasks, pairs) in an 8-segment example. Reduction to 7 tasks can be achieved if triples are included in the mix. We also provide a theoretical framework for further task reduction by applying a hierarchical latent class tree (LCT) structure to reduce segment similarity. Preliminary results with and without adjustment for scale confounds suggest that the LCT approach not only yields further task reduction but also provides more meaningful segments. These methods can be implemented with commercial software such as Latent GOLD® and CHi-squared Automatic Interaction Detection (CHAID).

INTRODUCTION

Once a meaningful set of segments is obtained, it is a common practice to develop a typing tool to assign new cases to the most appropriate of these segments (Figure 1). Typing tools typically consist of a short battery of questions together with an algorithm for mapping each response pattern to the appropriate segment. To minimize respondent fatigue, the tool should be as simple as possible but not so simple that it fails to achieve acceptable classification accuracy.

Figure 1. A typing tool attempts to assign a new respondent into the most appropriate segment.



Typing tools can be *static*, with the same questions administered to all respondents, or *adaptive*, where the tasks administered to a given respondent depend upon their responses to previous questions. Magidson and Bennett (2016) described an approach to develop a *static* typing tool with simple paired comparisons which achieves high accuracy of classifying new cases into the correct latent class (LC) segment.

In Part 1 of this paper we begin by reviewing that static approach using MaxDiff data from a sample of 200 respondents. We then modify the approach to develop an adaptive tool and

compare the expected accuracy (and the number of tasks required to achieve 80% accuracy) for both approaches. We also compare and contrast our typing tool development approach to earlier approaches (e.g., Orme and Johnson, 2009; Lyon, 2016; Komendant, 2016).

Part 2 of this paper introduces a new hierarchical LC tree-based framework to reduce segment similarity and make it easier to interpret the segments. This framework facilitates further task reduction, as shown by applying it to the scale-adjusted latent class (SALC) model to achieve more meaningful segments. We also discuss differences in the typing tool tasks that might be expected after scale heterogeneity is removed from the segments. All analyses presented here were performed with Latent GOLD[®] and SI-CHAID[®].

PART 1: STATIC AND ADAPTIVE TYPING TOOLS

There are three key components to our typing tool development based on LC segments. The only change to make the static approach adaptive is to modify component 3.

1. Use latent class modeling to obtain segment-specific *worth* parameters (utilities) that define each segment.
2. Simulate responses to all potential typing tool tasks for a large sample generated from each segment population.¹
3. Use stepwise multinomial logit (MNL) model for the static approach, or CHi-squared Automatic Interaction Detection (CHAID) for the adaptive approach, to select the tasks to include in the typing tool.

Component 1: Latent Class Segmentation

The starting point is to obtain segment-level MaxDiff utilities which serve to define the LC segments. These utilities are often obtained using a tandem approach, where hierarchical Bayes (HB) is used in step 1 to obtain individual level utilities, and then LC is used to cluster these utilities in step 2. However, that approach is theoretically inconsistent² and typically yields segments that are less interpretable than segments obtained in 1-step by applying an LC choice model directly to MaxDiff responses. For that reason, we agree with Eagle (2013) and Magidson (2003) in recommending that segment-level utilities be obtained using a 1-step approach by applying a LC choice model directly to the MaxDiff responses³. The 1-step vs. tandem approaches are compared further in Appendix B: One-Step vs. Tandem Approach to Obtain Segments from MaxDiff Data.

Component 2: Simulating Respondents from Each LC Segment

In Magidson and Bennett (2016), the most preferred alternative for all possible (36) paired comparison tasks were simulated for 1000 generated respondents from each of eight LC

¹ As described later, potential tasks include best only responses to pairwise comparisons (pairs), triples, quads, etc.

² The multivariate normal (MVN) prior (used in step 1) is tantamount to the assumption that segments do not exist (MVN random effects), while latent class clustering on the individual utilities obtained in step 1 assumes the existence of $K > 1$ segments (MVN random effects within classes). Since the distribution of individual utilities cannot be governed both by overall MVN (step 1) and mixture MVN (step 2 assumes MVN within each segment), this approach is inconsistent. Replacing the MVN prior with the MVN mixture prior in Step 1 (using covariates in the upper model) does not solve the problem and may even yield a worse segmentation if the covariates are not related to the behavioral segments. See Appendix A for a summary of three different LC choice models for 1-step analysis of MaxDiff responses.

³ Sawtooth Software also cautions users regarding this tandem approach and recommend the 1-step approach.

segments based on the utilities that define each segment.⁴ These simulated responses were then used as predictors of the (true) segments in a stepwise multinomial logit (MNL) model to determine the pairs most predictive of the segments, for inclusion in the typing tool.

The two primary benefits of using this simulation strategy with generated respondents are

1. The sample size ($N = 8000$) is sufficiently large so that the simulated responses contained the appropriate variability needed to obtain reliable results from the stepwise MNL.⁵ The typical sample size for a MaxDiff experiment (200–800 cases) is insufficient to achieve reliable results. In particular, Magidson and Bennett (2016) showed that the resulting accuracy from a typing tool where tasks were obtained by a MNL analysis of the original MaxDiff sample fell far short of the accuracy from a comparable tool where MNL variable selection was applied to a large number of generated respondents.
2. Since LC modeling is probabilistic in nature, respondents are assigned to the most likely class, and thus there is a non-zero probability that respondents are misclassified. Since the simulation approach generates respondents directly from the LC segment populations, their true class membership is known, and hence the dependent variable in the MNL represents true segment membership. In contrast, an MNL designed to predict the class assignments for the original MaxDiff respondents mistakenly assumes misclassification error is zero.

It should be noted that the utilities defining the LC segments were estimated using the sequential logit model as implemented in the Latent GOLD[®] program (Vermunt and Magidson, 2005). An alternative 1-step LC approach for analyzing MaxDiff responses, proposed originally by Louviere (1993) and implemented in the Sawtooth Software Latent class modeling program, cannot be used to simulate data without some complications.⁶

Component 3: Methodology for Typing Tool Task Selection

In Magidson and Bennett (2016), the best subsets of tasks were identified using stepwise MNL to predict segment membership based on simulated responses to the tasks. The resulting typing tool was “static” in the sense that all future respondents to the typing tool are administered identical questions. In this paper, we extend the methodology in two directions:

1. We extend the static typing tool to an adaptive (dynamic) tool where questions posed to each respondent differ according to their earlier responses, and
2. We allow triples, quads, etc. into the mix of potential tasks to include in the typing tool.

The key to accomplishing these extensions is to replace stepwise MNL with the CHAID decision tree technique.⁷ Using the same simulated data and generated respondents, we simply

⁴ Simulating data for an equal number ($N = 1000$) of cases from each of the eight segments represents a non-informative prior, which corresponds to the common situation where the typing tool is used on future populations that are not necessarily the same as the population from which the MaxDiff sample respondents were selected. In the situation that the future respondents to be typed are representative of the original MaxDiff population, unequal samples from each segment can be used to match the class sizes estimated by the LC analysis. Production of the simulated data is straightforward using the simulation capability in the Advanced/Syntax version of Latent GOLD[®].

⁵ An additional sample of 9000 respondents were generated in the same way and used as a validation sample to estimate the classification accuracy that would be achieved from the typing tool with different numbers of pairs.

⁶ Simulation is not straightforward with the 1-step approach based on the original Louviere (1993) MaxDiff algorithm implemented in the Sawtooth Software Latent class modeling program because data so simulated would contain some responses where the same alternative is selected as both best and worst (see Appendix A).

⁷ We used Statistical Innovations' SI-CHAID package.

substitute the CHAID algorithm for the stepwise MNL for task selection, as illustrated in detail in the example below, to yield an *adaptive* typing tool.

CHAID's tree structure provides a natural way to determine the most statistically significant task to include in the typing tool based on responses elicited from previous tasks. CHAID also deals with categorical predictors in a natural way, making the extension from pairs to triples, quads, etc. transparent.⁸

MaxDiff Data: Sydney Independent Transport Inquiry

To illustrate the approach and to compare the performance of the adaptive typing tool with the static tool, we utilize the MaxDiff response data from the Sydney Independent Transport Inquiry. More specifically, this is an example of a Case 1 Best-Worst Scaling study conducted in 2010 to determine how Sydney residents differ in their prioritization of nine short-term improvements to the transportation network (Table 1). For further details, see Louviere, Flynn, and Marley (2015).

Table 1. Sydney Independent Transport Inquiry

Object number	Object name (transportation improvement)
1	More frequent off-peak trains between major centers
2	Improved peak rail capacity
3	More frequent bus services on major routes
4	Extensions of light rail services
5	Integrated fares
6	Integrated ticketing
7	Real-time arrival information
8	New cycleways; more bike and scooter parking
9	Trains use green power

With 9 objects, there are a total of $\binom{9}{2} = 36$ object pairs that were considered as potential tasks for inclusion in the typing tool. The study used a Balanced Incomplete Block Design (BIBD) with each respondent being administered 12 sets of 3 options. The best and worst responses to task #1 (Table 2) yields preference orderings among each of the 3 paired comparisons, namely {2, 4}, {2, 8}, and {4, 8}.

Table 2. Example of the of the BIBD design for choice task #1 of 12.

Object number	Object name (transportation improvement)	Best	Worst
2	Improved peak rail capacity		
4	Extensions of light rail services		
8	New cycleways; more bike and scooter parking		

⁸ As pointed out by Lyon (2016), use of stepwise MNL is clear with dichotomous (and numeric) predictors, but the extension from pairs to triples, quads, etc. is not so straightforward. On the other hand, the CHAID algorithm is designed for categorical predictors and thus has no problem with extensions beyond pairs.

With 12 choice tasks each yielding orderings among 3 pairs, it follows that orderings are obtained for all 36 pairs in this efficient design ($12 \times 3 = 36$).

Results from the Static Typing Tool Development

The starting point in the development of the static typing tool developed in Magidson and Bennett (2016) of these data was to obtain the utility parameters from an 8-segment LC Best-worst model using the 1-step sequential logit modeling approach as implemented in Latent GOLD[®].⁹ These segment-specific utility parameters serve to define the 8 segments (see Table 3).

**Table 3. Traditional latent class modeling results for 8-classes
(from Latent Gold® tutorial 8A).**

	Utility parameters							
Object name	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
More frequent off-peak trains between major centers	1.2	1.0	0.4	-0.1	-0.3	-0.7	-0.1	-0.4
Improved peak rail capacity	1.3	3.1	0.8	3.6	0.7	2.2	1.1	1.3
More frequent bus services on major routes	1.4	1.8	0.2	2.6	0.6	2.2	0.8	0.6
Extensions of light rail services	0.5	-1.9	-1.2	-1.8	-0.3	-1.4	-2.4	1.3
Integrated fares	-0.5	0.4	0.3	-1.0	-0.3	-0.3	2.0	2.0
Integrated ticketing	-0.7	0.4	0.1	-1.2	-0.4	-0.4	2.7	2.6
Real-time arrival information	-0.8	-0.1	0.4	-1.9	-2.0	-1.1	-1.3	0.4
New cycleways; more bike and scooter parking	-1.3	-2.6	-1.6	-2.0	0.4	0.9	-1.9	-3.6
Trains use green power	-1.0	-2.2	0.6	1.7	1.6	-1.4	-0.8	-4.1
Standard deviation	1.1	1.9	0.8	2.1	1.0	1.4	1.8	2.4
Size	0.21	0.18	0.14	0.12	0.12	0.11	0.08	0.04

The next step was to simulate “best” responses to all possible pairs for 1,000 respondents generated from each of the 8 segments (the training data) based on the LC segment utilities given in Table 3. As mentioned above, generation of a large number ($N = 8,000$) of respondents is necessary to obtain sufficient variation in the responses. Figure 2 shows the simulated “best” responses for pairs $\{1, 2\}$, $\{1, 3\}$, \dots , $\{8, 9\}$ from the first 20 respondents, together with their true class membership (class#). For each of the eight segments, “best” responses to all 36 pairs were simulated.

⁹ For a detailed description of how these parameters were obtained, see Choice Tutorial 8A. www.statisticalinnovations.com/wp-content/uploads/LGChoice_tutorial_8A.pdf

Figure 2. Simulated “best” responses for all possible pairs (n = 36).

ID	Choice.12	Choice.13	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	Choice.89	Class#		
1	2	2	1	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	2	2	1	2	2	1	2	2	1	2	1	1	1	1	2	6	
2	2	2	1	1	1	2	1	1	2	2	1	2	1	1	2	2	1	1	1	2	2	1	2	2	1	1	1	1	1	1	1	1	1	2	2	4	
3	2	2	2	1	2	1	2	2	2	2	1	1	1	2	1	1	2	2	2	2	2	2	1	2	2	1	1	2	2	1	2	2	2	2	2	2	
4	1	2	1	1	1	2	2	1	1	2	1	2	1	1	1	2	2	1	1	1	2	2	1	2	2	2	2	1	2	2	1	2	1	2	2	5	
5	2	2	1	2	1	1	1	2	2	1	1	1	1	1	1	1	2	1	1	2	1	1	2	2	1	2	2	1	1	2	2	1	2	2	2	1	8
6	2	2	1	2	1	1	2	2	2	1	1	1	2	1	1	2	1	1	2	1	1	2	2	1	2	1	1	2	1	1	2	2	1	2	2	2	2
7	2	2	2	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	2	2	1	2	1	2	1	2	2	2	2	2	8	
8	2	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	2	2	1	1	1	1	1	1	1	2	6	
9	2	2	1	2	2	1	1	2	2	1	2	2	1	1	1	2	2	2	2	1	1	2	2	2	2	2	1	1	2	1	1	1	1	1	1	4	
10	2	2	1	1	2	1	1	1	1	1	2	1	1	1	1	1	2	2	1	1	1	2	2	2	1	2	1	1	1	1	1	1	1	1	2	4	
11	2	2	2	2	2	1	1	1	1	1	2	2	1	1	1	1	2	1	1	1	1	2	2	2	2	2	2	1	2	1	1	1	2	1	2	4	
12	1	1	1	1	1	1	1	1	1	2	1	2	1	1	1	1	1	1	2	1	2	1	1	1	1	2	2	1	2	2	1	2	1	2	1	1	
13	2	2	1	2	2	1	1	1	2	1	2	2	1	1	1	1	2	2	1	1	1	2	2	2	1	2	2	1	1	1	1	1	1	2	2	4	
14	2	2	1	1	2	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	1	2	2	1	2	2	1	2	1	2	8	
15	2	1	2	2	2	1	2	2	1	1	1	2	1	2	1	2	1	1	1	1	2	2	2	2	1	2	2	1	2	1	2	1	2	1	2	5	
16	2	1	2	2	2	2	1	1	1	1	1	1	1	1	1	2	2	2	2	1	2	2	2	2	1	2	2	1	1	1	1	2	2	2	1	5	
17	2	2	2	1	2	1	2	2	2	1	2	1	2	2	1	1	2	1	2	2	2	2	2	1	1	2	1	1	2	2	1	2	2	2	2	1	2
18	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	2	1	1	2	1	2	2	1	1		
19	2	2	1	1	2	2	1	1	2	1	2	2	2	1	2	1	1	1	2	1	1	1	2	2	2	1	2	1	1	1	2	2	1	1	2	5	
20	2	1	1	2	2	1	1	2	1	1	2	2	1	1	1	1	2	2	1	1	2	2	2	2	2	2	2	1	1	2	1	1	1	1	2	4	

Segment #5

Segment #5

The final step was to analyze these training data using stepwise multinomial logit modeling (MNL) to determine the K most predictive pairs to include in the typing tool. Specifically, Stepwise MNL predicted true class membership (“Class#” in Figure 2 denotes true class) as a function of the 36 pairs. Data on an additional 9,000 persons per segment were simulated and used to validate the accuracy of a typing tool based on K pairs. Table 4 shows the top four pairs obtained from the stepwise MNL.

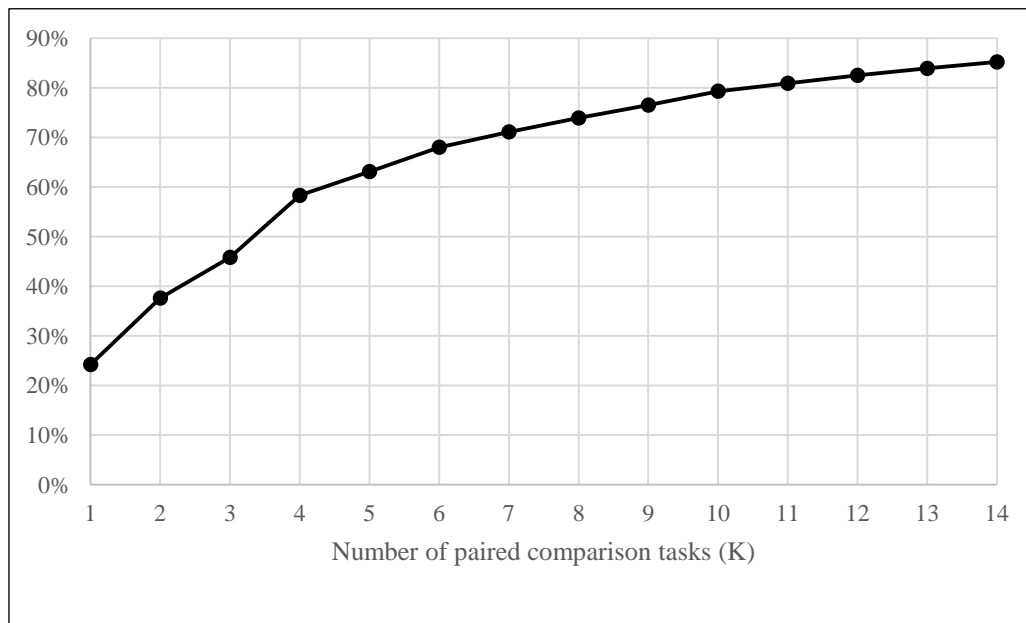
Table 4. The top four pairs used in the static typing tool.

Pair No./Object No.	Objects	Most Important
1 = (6, 9)	Integrated ticketing	
	Trains use green power	
2 = (7, 8)	Real-time arrival information	
	New cycleways; more bike and scooter parking	
3 = (4, 9)	Extensions of light rail services	
	Trains use green power	
4 = (3, 6)	More frequent bus services on major routes	
	Integrated ticketing	

Figure 3 summarizes the expected accuracy computed as a function of K , the number of tasks. These results show that a typing tool consisting of 8 pairs would be expected to reproduce the true segment membership with 74% accuracy.¹⁰

¹⁰ Accuracy is estimated based on the additional $N = 72,000$ cases generated as validation data, where task selection was based on $N = 8,000$ generated cases (training data).

Figure 3. Expected accuracy from a static typing tool with K paired comparison tasks.



We note that if MNL task selection were instead performed using the original sample ($N = 200$), the expected accuracy for a typing tool with 8 pairs would have been substantially lower (reduced from 74% to 63%), a result that supports the benefit of generating respondents.¹¹

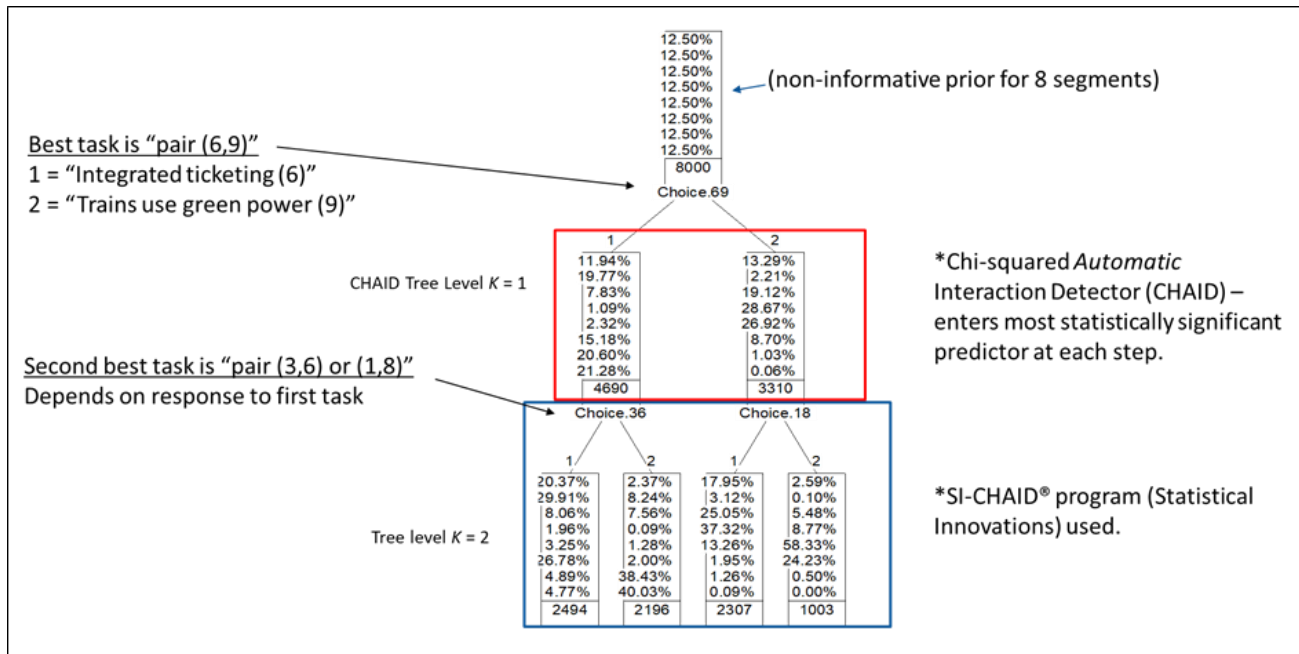
Comparison of Results: Static vs. Adaptive Typing Tool

For development of the *adaptive* tool the tree-based CHi-squared Automatic Interaction Detection (CHAID) algorithm replaces stepwise MNL for task selection. At each respondent subsample represented by a tree node, CHAID selects the predictor that is most significant (lowest p -value) based on a chi-squared test of each 2-way table of the predictor tabulated against segment membership. The sample is then split into subsamples according to the simulated responses for the most significant of the predictors (Magidson, 1994).

For comparison with the static typing tool we begin by limiting the potential tasks to pairs, using the same data generated previously for the development of the static tool. Figure 4 shows that the single best predictor is pair {6, 9} which asks respondents to choose the improvement they prefer: “Integrated ticketing” (object 6) or “Trains use green power” (object 9). Respondents are then administered either pair {3, 6} or {1, 8} depending on their response to the first paired comparison {6, 9}.

¹¹ Stepwise MNL on the original sample data with modal assignments as the dependent variable was found to overfit the data substantially. The in-sample accuracy with 8 binary predictors was 74% but fell-off to 63% when applied to the simulated validation data. The in-sample accuracy with 12 paired comparisons was 91.5% which fell-off to 69% on the validation data. For further details see Magidson and Bennett (2016).

Figure 4. Results for adaptive typing tool using pairs in CHAID.



The pair {6, 9} selected by CHAID as the most significant predictor overall, was also identified as most significant by the MNL analysis used to develop the static typing tool (recall Table 4). As shown in Figure 4, CHAID splits the training sample of 8,000 simulated respondents into two subgroups ($N_1 = 4690$ and $N_2 = 3310$) according to whether they chose “Integrated ticketing” (object 6) or “Trains use green power” (object 9) as most important. CHAID next selects either the pair {3, 6} or {1, 8} as the second pairwise comparison task for the adaptive typing tool depending on the subgroup (i.e., depending on their response to the first task {6, 9}). This dynamic feature of CHAID is what makes the typing tool *adaptive*.

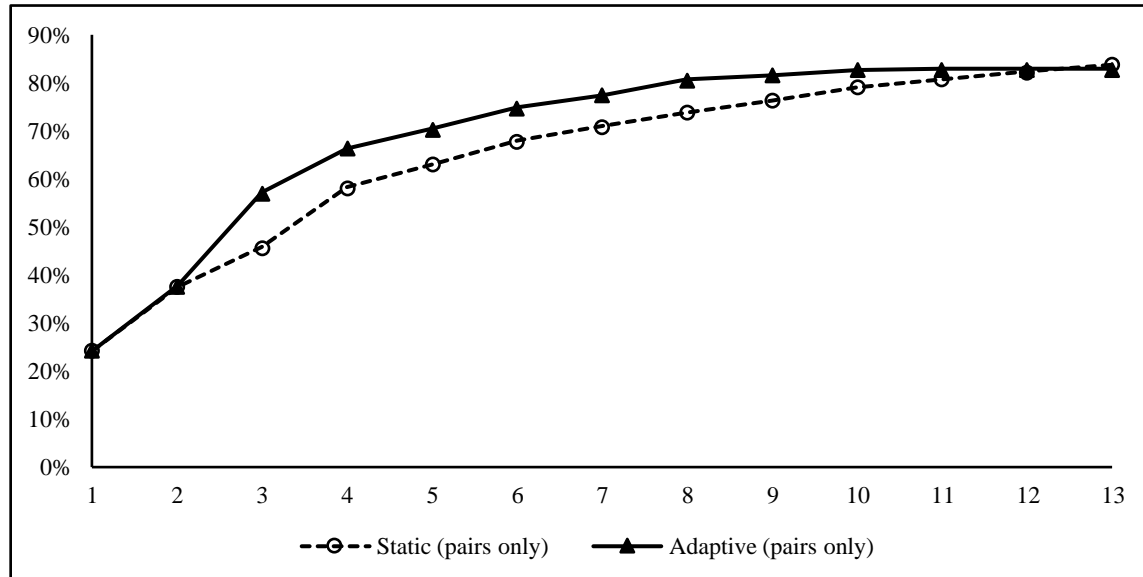
The end result of this CHAID tree is that each respondent is placed in one of the four terminal nodes (buckets) depicted at the bottom (“Tree level $K = 2$ ”) of Figure 4. For example, the first bucket consists of the 2,494 respondents who select statement “6” as best among the pair {6, 9} at tree level 1 and then select statement “3” as best among the pair {3, 6} at tree level 2. Since these respondents were generated, we know that 20.47% belong to true segment 1 and 29.91% to true segment 2, etc., as depicted in the first bucket (first terminal node) of the CHAID tree.

Depending on which of these four buckets one falls into, respondents are then assigned to the segment having the highest probability. Thus, persons in bucket #1 are assigned to segment 2, since the highest percentage for that bucket is 29.91% associated with segment 2. Persons in bucket #2 are assigned to segment 8 (the associated percentage being 40.03%), persons in bucket #3 are assigned to segment 4 (with percentage 37.32%) and those in bucket #4 are assigned to segment 5 (with 58.33% correctly assigned to segment 5).

Overall, only 38% would be assigned to the correct segment based on this hypothetical adaptive typing tool consisting of two pairwise comparisons per respondent, the same accuracy achievable by the static typing tool with two pairs. Figure 5 shows that for $K = 3$ or more pairs,

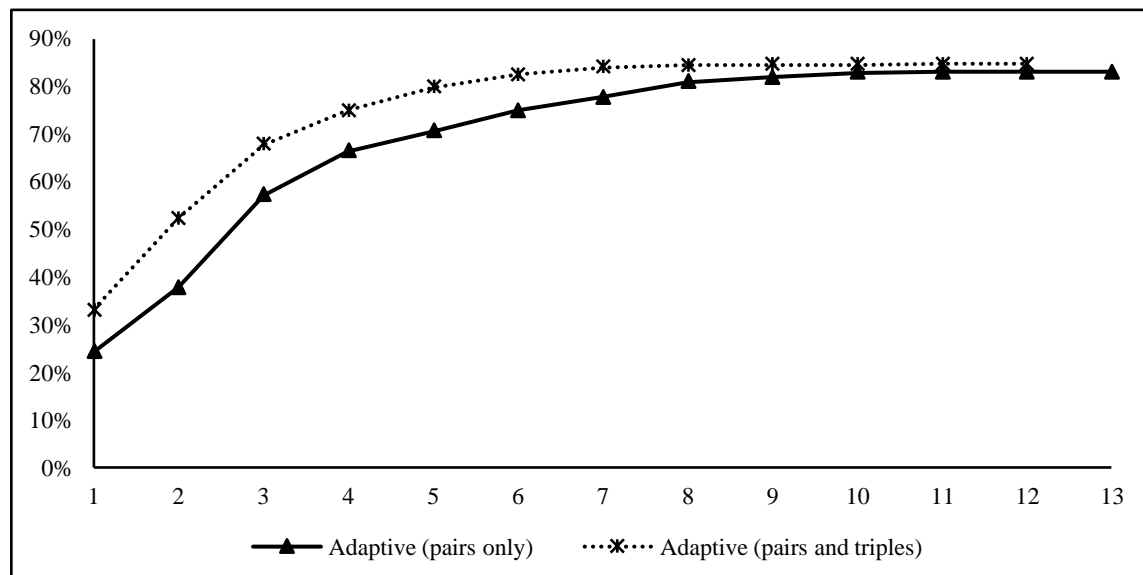
the adaptive tool yields higher accuracy than the static tool using the same number of pairs.¹² In particular, with 8 pairs, the adaptive approach yields 80% accuracy compared to 74% for the static tool.

Figure 5. Accuracy comparison of static and adaptive typing tools using K pairs.



The simulated response data used to develop the static typing tool was then expanded by simulating the “best” response to all possible triples and adding that to the mix for the adaptive typing tool development.

Figure 6. Accuracy comparison for adaptive pairs vs. adaptive pairs plus triples.



¹² The accuracy for both the static and adaptive tools was computed using the validation data of N = 72,000 simulated respondents. For the adaptive tool with (at most) K = 8 pairs, approximately 30% of the respondents were associated with terminal CHAID nodes defined by fewer than 8 pairs: K = 7 (13%), K = 6 (16%) and K = 5 (0.5%).

When the $\binom{9}{3} = 84$ triples were added to the mix, the triples tended to be selected ahead of pairs by CHAID and overall accuracy improved, especially for $K < 8$ tasks. In particular 80% accuracy was achieved in the adaptive tool (Figure 6) with at most 5 tasks (pairs and/or triples) per respondent, as compared to $K = 8$ tasks with pairs only.¹³

Summary of Results: Static vs. Adaptive

- Limited to pairs, the CHAID tree stopped after 13 levels, achieving accuracy of 85% (84% in validation data), the same level of accuracy achieved by the static approach where all respondents reply to 13 paired comparisons. In contrast, the *average* number of pairs to which respondents were exposed with the adaptive approach was 8, with some being classified after responding to as few as 5 pairs. Thus, the adaptive tool reduces fatigue, with no respondent required to reply to more than 13 pairs, and most only needing to reply to 8 or fewer pairs.
- With triples allowed to enter the mix, CHAID selected triples over pairs for the initial tree splits, but then completed the tree with binary splits, terminating after 12 levels, and yielding validation accuracy of 85%.¹⁴ Thus, including triples improves the accuracy slightly and also reduces the fatigue somewhat, the average number of tasks to which respondents are exposed being 7.

We conclude that a major advantage of adaptive typing tools is that fewer tasks are required in order to achieve high accuracy. In this example with 8 segments, 13 tasks are needed to achieve accuracy in the range of 83% to 85%. In contrast, using the adaptive approach, that number is reduced to an average of 8 tasks.

Table 5. Comparison of expected classification accuracy for static and adaptive approaches.

	Static	Adaptive	
	MNL regression	Pairs Only	Pairs and Triples
Accuracy-Validation	83.9%	83.0%	84.6%
Accuracy- Sample	84.8%	88.7%	93.1%
Average no. of items	13.0	8.3	7.0

If choice tasks with three objects (triples) are also allowed in the experiment, along with pairs, that number of tasks is further reduced to seven. It is also possible that the inclusion of quad-based tasks could result in further task reduction, although that hypothesis has not been formally tested at this point. It is clear that the results suggest that adaptive surveys have the potential to improve both the classification accuracy and efficiency in choice experiments. Adaptive surveys are, however, more difficult to program. The algorithms needed to determine the “best task” at every choice juncture are not typically available “off the shelf.” Researchers and policymakers will need to decide whether investing in this type of capacity is necessary for their choice experiments.

¹³ Note: the overall “best” triple was the one containing objects “6-8-9” for comparison.

¹⁴ Since MaxDiff tasks can be represented by selection of a triple followed by response to the relevant paired comparison MaxDiff tasks are included in the accuracy comparisons. MaxDiff tasks were never selected by CHAID (see Appendix D).

Comparison with Other Approaches

Typing tool development for segments derived from MaxDiff Response Data has been addressed earlier by Orme and Johnson (2009), Lyon (2016) and Komendant (2016), which we collectively refer to as Naïve Bayes Classifier (NBC) approaches. Our approach differs fundamentally from these approaches. Specifically, our approach:

- yields simpler typing tools (by excluding complete MaxDiff tasks and asking only for the “best” responses),¹⁵
- defines segments in terms of their preferences (i.e., parameters), not in terms of those respondents assigned to the segments, and
- computes expected accuracy directly, not as the hit rates to the assigned classifications for the original MaxDiff respondents.

One-Step vs. Tandem Approach to Obtain Segments from MaxDiff Data

Our approach requires only that segment-level MaxDiff utilities are available from some process. We recommend that the sequential logit (also known as “best-worst”) model, be used as the process to obtain these utilities, directly from responses observed from the original MaxDiff respondents (see Appendix A). These estimated utilities serve to define each segment. In contrast, a tandem approach is commonly used where individual level utilities from a hierarchical Bayesian (HB) analysis of the MaxDiff responses are obtained as a first step, and segments are then obtained by clustering these utilities as a second step.

Presumably, persons who use HB may choose the tandem (two-step) approach because they desire to obtain segments that are most consistent with the HB utilities. However, our research suggests the surprising result that the 1-step LC approach in fact yields segments that are not only more meaningful but also *more* in agreement with HB utilities than segments obtained using the tandem approach (see Appendix B for an example with the data used in Lyon, 2016).

Thus, regardless of the availability of HB utilities, we follow Magidson (2003) in recommending that the 1-step approach be used to obtain the segment-level utilities. Moreover, if HB utilities are available, the resulting segment-level utilities obtained using the 1-step approach may provide additional insight into the heterogeneity that exists among these HB utilities (for an example of this, see Magidson, 2018).

PART 2: FURTHER TASK REDUCTION WITH STRUCTURED LC MODELING

The accuracy of a typing tool is determined in part by segment differences. All things being equal, the more different the segments, the higher the accuracy that can be achieved by the resulting typing tool. In this section we consider ways to improve the typing tool accuracy by refining the segments themselves to be more different in meaningful ways.

Segments that are similar with respect to their preferences, by definition, are more difficult to differentiate. Reducing the number of segments by combining similar segments, increases accuracy because there are fewer segments into which new respondents can be misclassified. Therefore, if it is possible to reduce the number of segments (from the eight in the standard LC solution, in our example) while retaining the core segment differences, it should be possible to

¹⁵ Our research suggests that requesting the worst choice adds little value to the typing tool. See Appendix D.

form segments that are both more useful (from a practical substantive standpoint) and more accurate (from a statistical classification standpoint). A related benefit is that the resulting typing tool can achieve high accuracy with fewer tasks. The easiest way to reduce segment similarity is to identify and combine similar segments, each segment grouping having preferences that differ from the other segment groupings.

Part 2 of this paper addresses the following questions:

- Can the standard LC modeling paradigm be modified in a statistically sound manner to yield segments that are more policy-relevant?
- With a relatively small number of segments that show clear differences in preferences, can a typing tool achieve high accuracy with only a few tasks?
- What is the least number of tasks that yield high accuracy?

Goal #1: Identifying a Smaller Subset of Policy-Relevant Segments

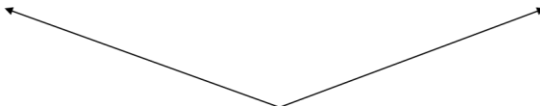
Results from standard LC modeling suggest that at least eight segments are needed to capture the different preference groups for these survey respondents (recall Table 3). Because eight is a relatively large number of segments, interpretation presents a challenge to policymakers and managers who look to focus on a smaller number of core “themes” that reflect *primary* segment differences. Transportation planners would be hard-pressed to implement the changes to the city’s system where eight distinctly different choice preferences exist. If it were possible to reduce the eight segments down to a smaller number that reflected the most salient differences among the survey participants, then there is a greater opportunity to make meaningful improvements.

The standard LC modeling paradigm relies on information criteria such as the Bayesian Information Criteria (BIC) to determine the number of classes. For the Sydney transport data, Table 6 shows that whether we use a standard LC model or a scale-adjusted (SALC) model (to provide segments that are more meaningful), at least 8-classes would be suggested by the BIC. (The LC solution with the lowest BIC value is preferred.)¹⁶

¹⁶ For an introduction to SALC models and explanations as to how they make segments more meaningful see Appendix C.

Table 6. BIC Comparison for standard LC models and SALC models with 2 scale classes.

Standard LC Model				SALC Model			
# classes	LL	BIC	Npar	# classes	LL	BIC	Npar
1	-3704.4	7451.2	8	1	-3643.1	7339.2	10
2	-3576.5	7243.0	17	2	-3525.8	7152.3	19
3	-3479.1	4096.0	26	3	-3439.2	7026.8	28
4	3419.5	5024.4	35	4	-3386.7	6969.5	37
5	-3366.1	6965.2	44	5	-3331.9	6907.6	46
6	-3320.5	6921.8	53	6	-3293.6	6878.7	55
7	-3285.3	6899.0	62	7	-3260.2	6859.4	64
8	-3252.5	6881.2	71	8	-3228.4	6843.7	73



In both models, the BIC falls as more classes are added, suggesting at least 8 classes

Policymakers using latent class analysis to identify segments in their data are often faced with the need to reduce the number of classes obtained in standard solutions to a smaller set of more *policy or marketing relevant* segments. This has not been an easy problem to solve largely because the criteria used to inform decisions about segment extraction, the BIC and similar fit statistics, are sensitive to *any* class differences, not just *primary* differences that correspond to general core themes. Since traditional fit statistics do not differentiate between primary and secondary segments, it is necessary to consider modeling approaches capable of accounting for a hierarchical structure. Before trying the more formal structured LC modeling approaches, we begin in a qualitative way to achieve segment reduction and show how this can lead to a typing tool with higher accuracy.

A qualitative inspection of Table 7 suggests the following:

- Segments 1 and 2 share a preference for “More frequent off-peak trains.”
- Segments 7 and 8 share a preference for “Integrated fares” and “Integrated ticketing.”
- Segments 3, 4, and 5 share a preference for “Trains use green power.”

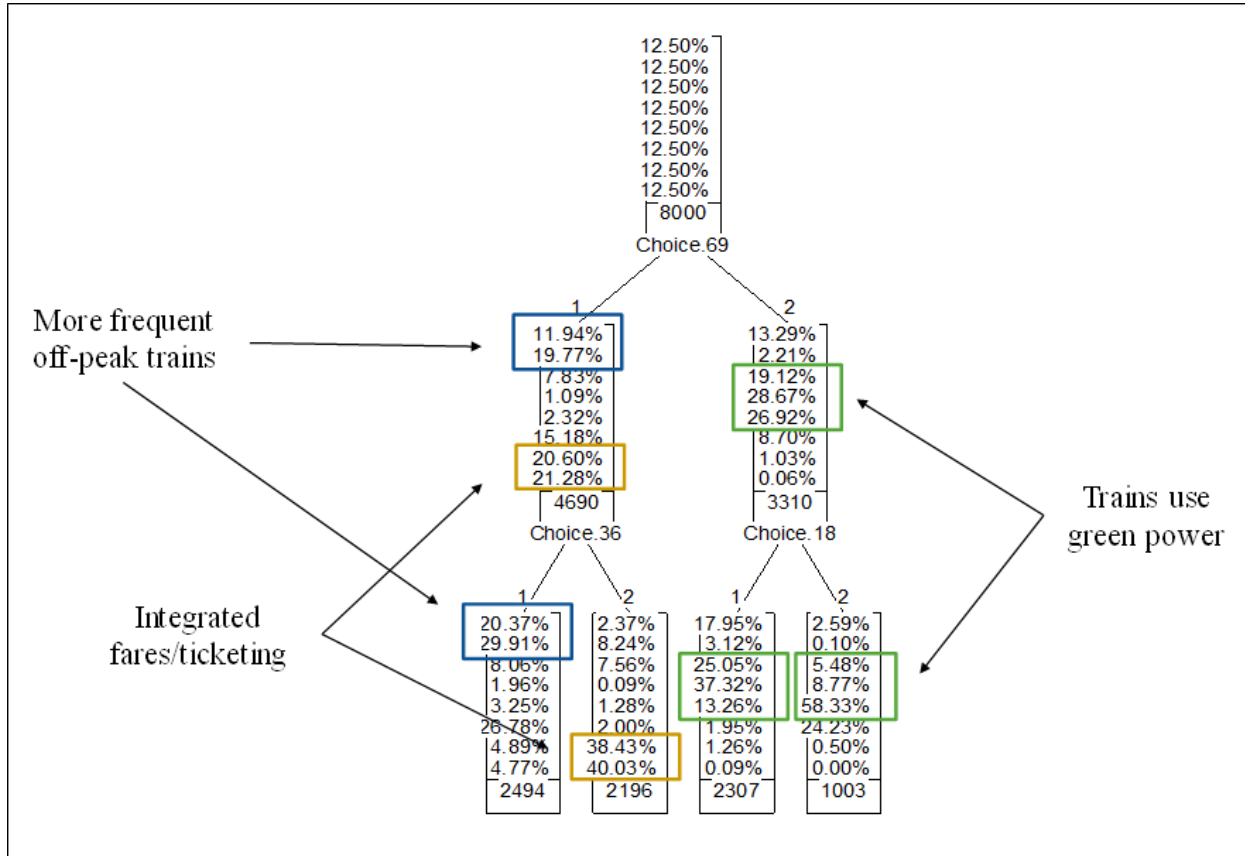
Table 7. Identifying similar segments among the 8-class solution.

Object name	Utility parameters							
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
More frequent off-peak trains between major centers	1.2	1.0	0.4	-0.1	-0.3	-0.7	-0.1	-0.4
Improved peak rail capacity	1.3	3.1	0.8	3.6	0.7	2.2	1.1	1.3
More frequent bus services on major routes	1.4	1.8	0.2	2.6	0.6	2.2	0.8	0.6
Extensions of light rail services	0.5	-1.9	-1.2	-1.8	-0.3	-1.4	-2.4	1.3
Integrated fares	-0.5	0.4	0.3	-1.0	-0.3	-0.3	2.0	2.0
Integrated ticketing	-0.7	0.4	0.1	-1.2	-0.4	-0.4	2.7	2.6
Real-time arrival information	-0.8	-0.1	0.4	-1.9	-2.0	-1.1	-1.3	0.4
New cycleways; more bike and scooter parking	-1.3	-2.6	-1.6	-2.0	0.4	0.9	-1.9	-3.6
Trains use green power	-1.0	-2.2	0.6	1.7	1.6	-1.4	-0.8	-4.1
Standard deviation	1.1	1.9	0.8	2.1	1.0	1.4	1.8	2.4
Size	0.21	0.18	0.14	0.12	0.12	0.11	0.08	0.04

Goal #2: Improving Segmentation Accuracy in the Sydney Transportation Inquiry

In addition to improving the meaning of segments, segment reduction also results in increased classification accuracy. To illustrate this reduction, suppose that some of the original eight segments are combined—the boxes in Figure 7 indicate the segments that are combined (i.e., classes 1 and 2 are combined; classes 3, 4 and 5 are combined; and classes 7 and 8 are combined). Revisiting our simple adaptive typing tool example based on CHAID with only two paired comparisons (recall Figure 4), the computations below (Figure 7) illustrate the increased accuracy expected if these similar segments were combined.

**Figure 7. Increased accuracy resulting from combining similar segments:
Illustration with adaptive typing tool based on 2 pairs.**



Specifically, combining the segments in the rectangles shown in Figure 7 increases accuracy from 38% to 68%. For example, the first bucket, consisting of 2,494 respondents, were assigned originally to segment 2 since the segment 2 percentage (29.91%) is the highest among the 8 segments. So, about 30% (29.91%) of these 2,494 respondents would be classified correctly. Overall, with 8 segments, we see that 38% would be classified correctly:

$$\frac{((0.30 * 2494) + (0.40 * 2196) + (0.37 * 2307) + (0.58 * 1003))}{8000} = 38\%$$

Alternatively, if true segments 1 and 2 were combined into a single joint segment, a total of 20.37% + 29.91% = 50.28% of these 2,494 respondents would be classified correctly into this combined segment, and overall, the accuracy increases to 68%:

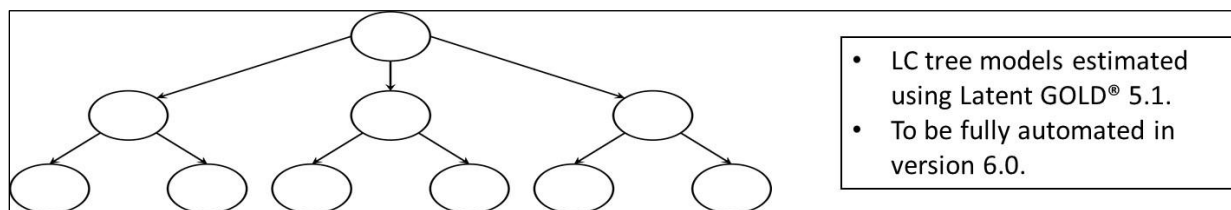
$$\frac{((0.20 + 0.30) * 2494 + (0.38 + 0.40) * 2196 + (0.25 + 0.37 + 0.13) * 2307 + (0.05 + 0.09 + 0.58) * 1003)}{8000} = 68\%$$

So, the question becomes one of how to justify reducing the number of segments if the BIC suggests we need 8 segments to explain all of the heterogeneity in the data. We will see that the answer to this question is to replace the standard *unstructured* LC modeling paradigm with the hierarchical latent class tree (LCT) structure proposed by van den Bergh et al. (2018).

Goal #3: Formalizing (Tree) Structure in Latent Class Analysis

The development of LCT was motivated by the observation that in practice, the number of policy relevant segments for strategic purposes is often 3 or 4, each of which represents a different theme. Figure 8 depicts a hierarchical LCT structure that first identifies three *core* or *theme* classes, respondents in each class differing in their primary preferences. Each of these *primary segments* splits further into two subsegments (*child* classes) to reveal secondary differences. The entire process, displayed in Figure 8, results in a total of $3 \times 2 = 6$ terminal segments at the bottom of the tree. The splitting process would continue further if warranted—i.e., if additional secondary differences were found to be statistically significant.¹⁷

Figure 8. LC Tree Models provide an alternative structure to standard LC models.



The identification of central *root*, *theme*, or *basic* level classes represents arguably the most important step in the LCT approach. The importance stems from the fact that these classes reveal the *primary* differences in preferences, and each of these themes are often maintained in all subsequent splits of that theme class.¹⁸ Since the primary differences are generally meaningful from a policy perspective, the LCT paradigm provides the interpretative power needed to make sense of the segments. In this respect LCT improves over the standard LC paradigm which often results in many unstructured classes which are often difficult to interpret in a meaningful way. In summary, the LCT paradigm explains *all* the heterogeneity in the data by applying a hierarchical tree structure that begins at the root of the tree with a small number of root (theme) classes.

To help identify the value K^* representing the number of theme classes, van den Bergh et al. (2018) proposed a new statistic, called the relative log-likelihood (RLL). This statistic summarizes *incremental* improvements in the log-likelihood (LL) as additional classes are extracted.

The RLL statistic is

$$RLL_{K,K+1} = \frac{\log L_{K+1} - \log L_K}{\log L_2 - \log L_1}$$

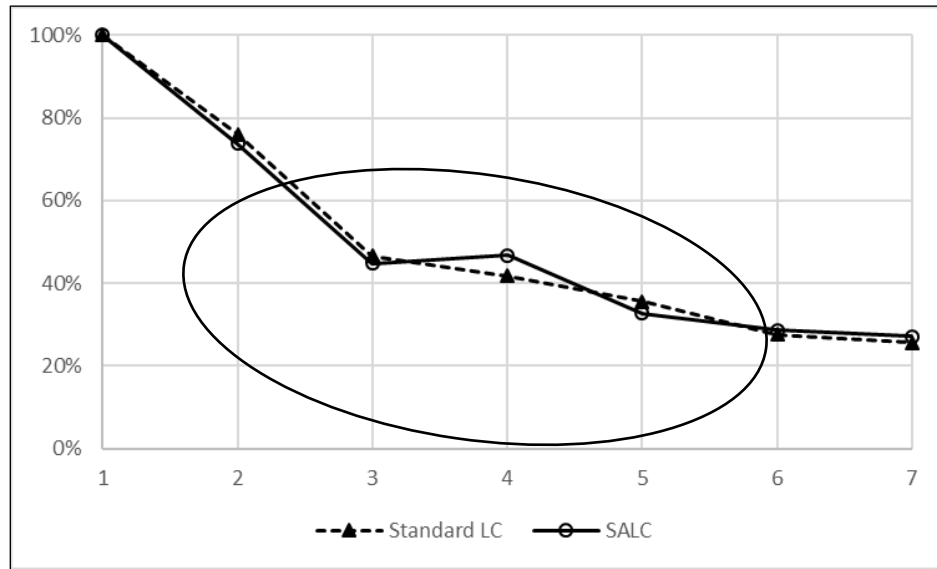
RLL recognizes that the largest increase in the log likelihood (LogL) occurs when the number of classes increases from one to two ($\log L_2 - \log L_1$).¹⁹ As a result, increases in $\log L$ that occur for $K > 2$ classes are expressed relative to this initial increase. The approach to identify the K^* is analogous to the use of the scree plot in factor analysis to determine the number of factors. Using the Sydney Transportation Inquiry as an example, the scree plot in Figure 9 shows that the relative change in LogL levels off beginning at $K = 3$, so we select 3 as the number of theme classes.

¹⁷ While the LCT procedure can be employed using Latent GOLD 5.1, it will be fully automated in release 6.0.

¹⁸ See Table C2 in Appendix C for an example where the theme from class 1 is maintained.

¹⁹ The 1-class model is known as the *aggregate* model which assumes a single homogeneous population.

Figure 9. Relative log-likelihood (RLL) by K = number of classes.



The general idea is that while adding another class beyond class $K^* = 3$ improves the model fit somewhat, if the improvement “levels off” when adding class K^*+1 , incremental improvement after $K^* = 3$ theme classes is treated as “secondary” to the earlier “primary” differences revealed by the first K^* segments. In summary, while eight segments were selected under the standard LC paradigm, the LCT paradigm yields three *core* segments that identify *primary* preference differences among the MaxDiff survey respondents.

Relating these theme classes to the objects (attributes), Table 8 shows that a defining characteristic differentiating theme class 1 from the others is that they have a high utility (0.9) for “More frequent off-peak trains,” and thus are more likely to choose this option as best (i.e., better than the average object) than respondents in the other classes. Similarly, theme class 2 respondents are more likely to choose the option “Trains use green power” and theme class 3 tends to prefer “Integrated fares” and “Integrated ticketing.” The shading in Table 8 summarizes the defining characteristics and associated utilities distinguishing the 3 classes.²⁰

Defining characteristics of the theme classes are highlighted in Table 8 along with the associated utilities. Despite the obvious differences between the classes, the statistics in the bottom two rows of Table 8 suggest that membership in theme class #2 may be confounded with scale.²¹

²⁰ Note that the standard deviation column in Table 8 also flags these objects as most important in explaining the heterogeneity (highest standard deviation). See Appendix B for the use of these standard deviations in practice and how they relate to the standard deviations of the individual utilities derived from HB.

²¹ Specifically, the relatively small standard deviation (0.75) for the utilities defining theme class #2 suggests that this class not only contains those who prefer green power but also includes (low scale) respondents who tend to be somewhat inconsistent in their responses, even if their preferences tend to be more in line with class 1 or class 3. That is, low scale respondents may be included in class 2 simply because this class is the one that tends to have estimated utilities of lower magnitude. See Table C1 in Appendix C for the SALC model alternative to Table 8 where the scale confound is removed, transforming the utility parameters to preference parameters.

Table 8. Utility parameters for the three “theme class” model.

	Utility parameters			
Object name	Class 1	Class 2	Class 3	Std. Dev.
More frequent off-peak trains between major centers	0.9	-0.1	0.0	0.46
Improved peak rail capacity	2.2	1.0	1.0	0.58
More frequent bus services on major routes	1.6	0.9	0.4	0.47
Extensions of light rail services	-0.8	-0.6	-1.2	0.20
Integrated fares	-0.3	-0.3	1.3	0.63
Integrated ticketing	-0.4	-0.4	1.6	0.82
Real-time arrival information	-0.6	-1.1	-0.3	0.31
New cycleways; more bike and scooter parking	-1.4	-0.4	-1.9	0.61
Trains use green power	-1.3	0.9	-1.0	1.01
Standard deviation	1.28	0.75	1.22	
Size	0.42	0.38	0.20	

IMPLICATIONS FOR A TREE STRUCTURED ADAPTIVE APPROACH

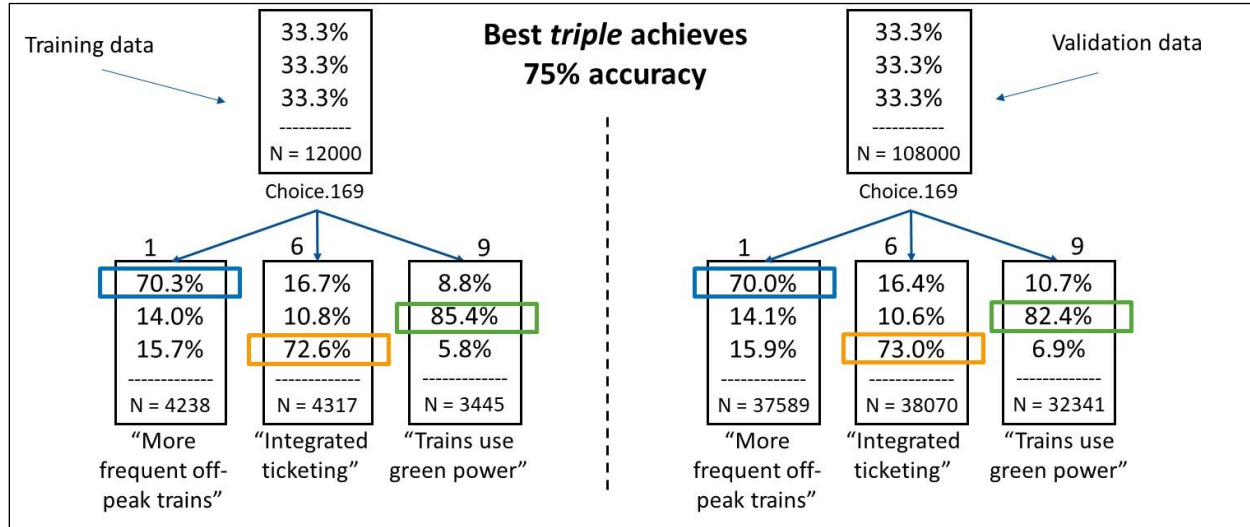
The basic idea is that an adaptive typing tool can take advantage of the LC Tree structure to achieve improved parsimony by identifying fewer “golden” questions (tasks to be included in the typing tool). In the extreme, we consider the possibility of including only *a single golden question* for each level of the LC tree. In the current example, at the first (core) level, where there are 3 “theme classes,” we might expect that a triple would achieve higher accuracy than a pair. At the second level, where each theme class may be split into 2 child classes, we might expect that a pair achieves high accuracy. The end result would be a substantial further reduction to only two tasks and thus make the classification of new respondents more economical.

Thus, a typing tool can first classify respondents into the most appropriate *theme* class, and then further refine the classification of these segments into the most appropriate second level segment using additional tasks. This approach has the additional advantage that persons who terminate the survey after the first task would still provide sufficient information for level-1 classification.

Since the theme classes should reflect differences in respondents’ *primary preferences* free from differences in *scale* we employ the SALC model with the current data so that the resulting class-specific parameter estimates are more clearly interpretable as preferences. (Compare the utility parameter estimates in Table 8 with the corresponding *preference* parameters shown in Table C1 of Appendix C.) The resulting SALC Tree splits each of these theme classes into 2 child nodes to reveal *secondary preferences* as depicted in Figure 8. Again, these secondary preferences should be relatively free from scale differences. (See Table C2 for the resulting preference parameters associated with the 2 child nodes of theme class 1.)

Based on data simulated from this SALC model²², Figure 10 shows that with one task consisting of a choice among the three options (1, 6, and 9), on average, respondents can be classified with 75% accuracy into the most appropriate theme class.

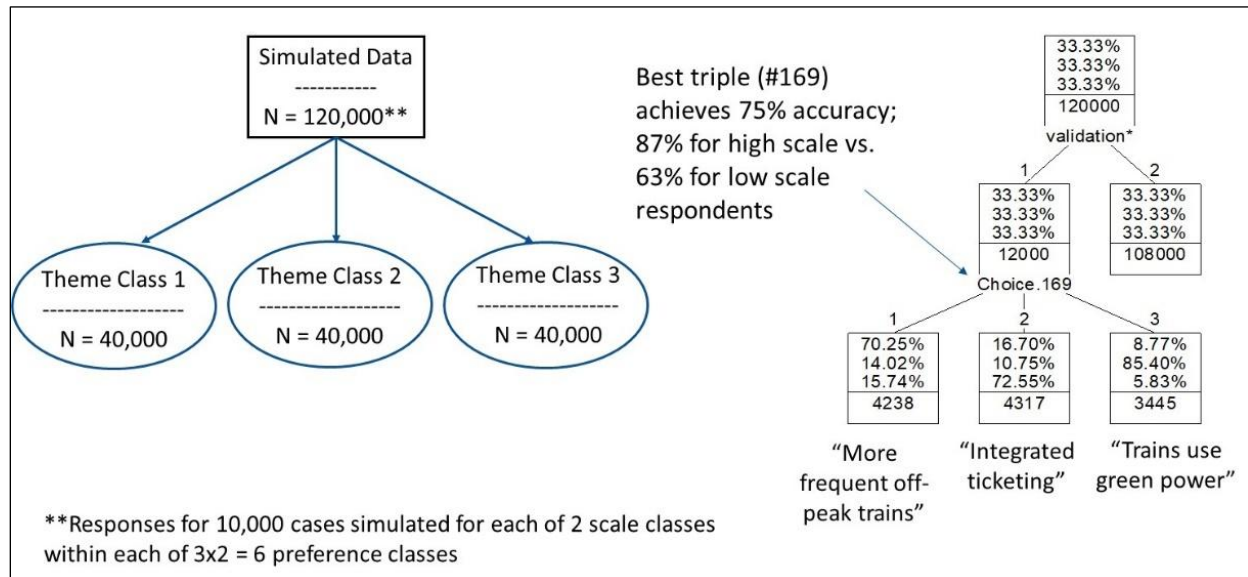
Figure 10. Accuracy using one 3-object task (triple) for classification at the theme level.



A more complete picture of the data generation and the resulting CHAID analysis, is provided in Figure 11. The graphic on the left shows the sample sizes for each of the three theme classes. Ten percent of the 40,000 respondents generated into each theme class were then randomly split into a 10% training sample and a validation sample (90%). The CHAID tree on the right shows the first step of the analysis on the training data (validation = "1"), where the triple {1, 6, 9} is selected as the most significant predictor and used to assign respondents to the appropriate theme class.

²² Since each of the 3 SALC model theme classes split into 2 child classes, the simulation assumed that the size of each of these 3x2=6 segments was identical, and that each consisted of an equal number of lower and higher scale respondents. Specifically, one thousand respondents were generated from each 6x2=12 segment x scale class populations for the training sample, and an additional 9000 respondents for the validation sample. In practice, whether or not to split into 2 child classes is determined by the BIC criteria. The 1-class and 2-class models are compared and if the 2-class model has a lower BIC, this theme class is split into 2 child classes. See van den Bergh et al. (2018) for details. Since we simulated based on the SALC model, we can estimate accuracy separately for the high and low scale respondents. Our simulation suggests that the accuracy obtained for high scale respondents would be 87% and for lower scale respondents, who may be less certain about their preferences, expected accuracy is 63%.

Figure 11. Data simulated from SALC tree model demonstrates lower accuracy for low scale respondents.



The classification accuracy could be improved by adding another task. Alternatively, the 1-task accuracy would be higher if the typing tool were only administered to high scale respondents. In particular, the 1-task classification accuracy improves from approximately 75% to 87% for high scale respondents, which means that fewer tasks would be required.²³

Use of an LCT-structured segmentation allows integration of the segmentation and typing tools in ways that were not possible previously. Specifically, we obtained three theme classes, and by examining the preference parameters we were able to identify defining characteristics for each theme class. In many cases, identifying such “defining characteristics” directly from the preference parameters may suggest the best tasks to use in a typing tool. In the current situation, we see that task {1, 6, 9} identified by CHAID might also be identified from the preference parameters (Table 8), since the objects 1, 6 and 9 are among the 4 objects highlighted.

By engaging in a “deeper dive” into the tree, the LCT approach allows further refinement of our understanding of respondent preferences. The CHAID analysis of data simulated from the SALC 3x2 tree model (with 3 theme classes each of which splits into 2 child nodes as depicted in Figure 8) shows an 81% accuracy would be expected at the 2nd level of classification (see Figure 13), given that one is classified correctly into the most appropriate theme class.

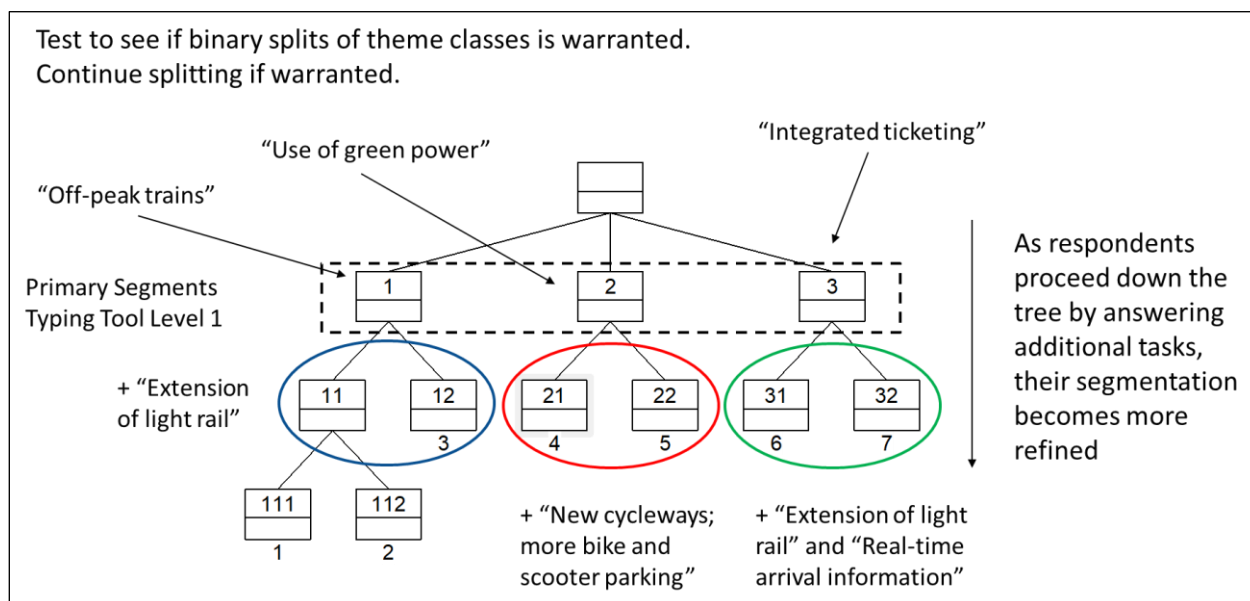
For example, Table C2 in Appendix C shows that sub-segments 11 and 12 share the primary theme class 1 preference for “More frequent off peak trains.” Examining the preferences of these sub-segments in more detail we find that they differ from each other in their preference for “Extension of light rail” (see Figure 12). Thus, we might expect that “Extension of light rail” might be paired with another object to serve as the best pair to classify theme class 1 respondents further into the most appropriate sub-segment.²⁴

²³ Typically, it is difficult to identify respondents by scale class in advance, but it is not impossible. If, for example, known covariates are predictive of scale class, it may be possible to use this information to predict the average number of tasks needed for classifying new cases.

²⁴ Based on this formulation of latent class analysis, a latent class tree model with no splits on the theme classes simply reduces to the standard latent class model. This suggests that standard latent class models are special cases of latent class tree models, having no significant secondary heterogeneity.

By using hierarchical trees, it is possible to administer tasks to respondents dynamically and segment them quickly. In addition, it is possible to conduct “deeper dives” on respondents to reveal their secondary preferences by administering additional tasks. In this transportation policy survey shown in Figure 12, the first task includes three objects: one about “off-peak trains,” another about the “use of green power,” and a third about “integrated ticketing.” Respondents are asked to select the best option among these three, and their choice provides what can be understood as the “primary preference” or “theme segment” assignment.

Figure 12. Create additional meaningful segments to explain more heterogeneity.

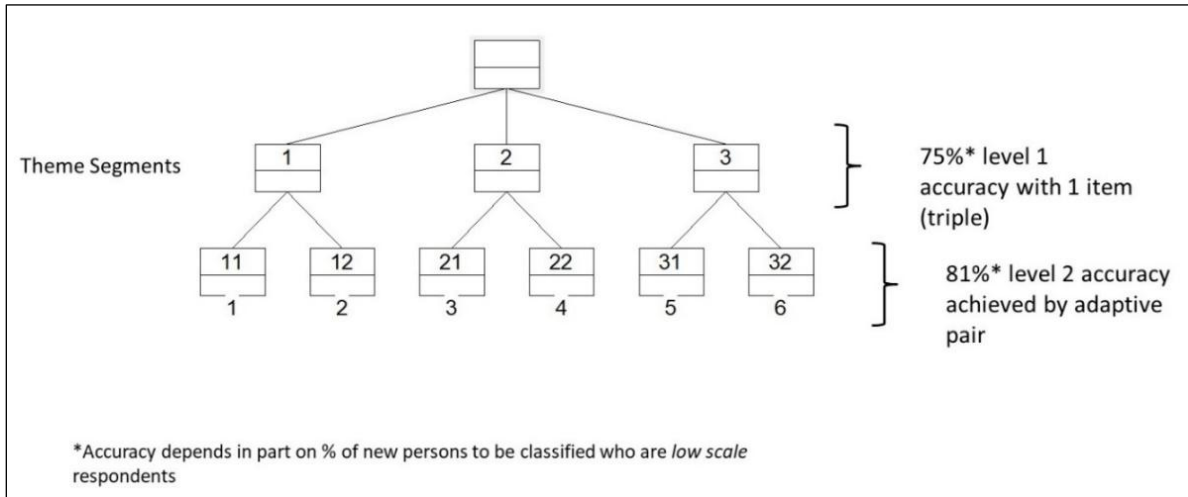


Preference at this root level of the LCT is often maintained with the respondent as respondents make additional choices.²⁵ For example, if a respondent identifies “off-peak trains” in the first task, they will continue to be identified as a person that prefers that policy. Secondary preferences are simply added to the primary preference allowing the segmentation to become more refined.²⁶

²⁵ Based on this formulation of latent class analysis, a latent class tree model with no splits on the theme classes simply reduces to the standard latent class model. This suggests that standard latent class models are special cases of latent class tree models, having no significant secondary heterogeneity.

²⁶ See Table C2 in Appendix C for an example of the resulting preference parameters for the 2 child class segments formed by splitting theme class 1.

Figure 13. Final 2-level hierarchical typing tool achieves high accuracy with 2 adaptive tasks.



DISCUSSION

Data from a typical MaxDiff exercise were used to illustrate the traditional approach to LC segmentation. According to the standard LC paradigm, BIC suggested that at least eight classes are needed to explain all the heterogeneity in the data. Using the resulting class-specific parameters to define the segments, following Magidson and Bennett (2016), a simulation approach with generated respondents was employed with stepwise multinomial logit modeling (MLM) to select a relatively small subset of tasks to be used in a static typing tool, where all respondents are administered the same tasks. It was shown how this simulation approach could be used to develop an adaptive tool with a reduced number of tasks by substituting the CHAID algorithm for MLM.

The simulation strategy was shown to be superior to earlier approaches for typing tool development, the two major benefits being:

1. A large sample size can be generated easily. The large sample is needed to prevent inflated estimates of accuracy (overfitting) resulting from analysis of data from a relatively small sample that is overly affected by sampling fluctuation.
2. The approach allows us to work directly with *true* segments, rather than *modal* assignments, thus eliminating artifacts and inflated accuracy associated with ignoring misclassification error.

In part 2 of this paper, we explored the use of a structured hierarchical tree-based paradigm for LC segmentation that allows integration of the segmentation and typing tools in ways that were not possible previously. In particular, replacing the standard LC paradigm with the LC Tree paradigm reduced the number of segments from eight to six. These six segments were structured in a way that identified three meaningful theme classes which formed the root of the tree, each of which differed in their *primary* preferences. Each of these theme classes then split into two sub-segments which formed the second level of the tree, revealing meaningful *secondary* differences within each of these theme classes.

By taking advantage of the meaningful hierarchical structure in these six segments an adaptive typing tool was developed which drastically reduced the number of tasks further, resulting in only two tasks. An additional advantage of employing the hierarchical tree structure is that if respondents terminate the survey early, say after answering only the first of the two golden questions, they can still be classified into the most appropriate theme class. Table 9 summarizes the advantages and disadvantages of the various methods of developing typing tools that were described here.

Table 9. Advantages and disadvantages of the 3 typing tool approaches.

Approach	Advantages	Disadvantages
Static	All participants answer the same survey questions/ easier to program	Respondents exposed to more items to achieve desired accuracy
Dynamic	Reduces respondent burden/ fewer items; improves accuracy	More difficult to program
Tree-Structured Dynamic	More meaningful segments for strategic purpose <ul style="list-style-type: none"> • More precise tool with only K tasks on average per respondent (here, K=2) • Early terminators can still be assigned to meaningful (K=1) 'theme' classes 	Even though analysis of MaxDiff data may yield meaningful structure, this is not guaranteed

The SALC Tree paradigm has been proposed as a replacement for the traditional LC paradigm (Magidson, 2018) and it has been shown that the resulting segments are actually more consistent with HB utilities than segments derived by clustering individual HB utilities. Research in this general area is ongoing.



Jay Magidson



John P. Madura

APPENDIX A: THREE DIFFERENT LC MODELS FOR ONE-STEP ANALYSIS OF MAXDIFF RESPONSES

There are primarily 3 different LC models that can be used to segment respondents based on their MaxDiff choices. All 3 approaches can be estimated with Latent GOLD®.

1. Sequential Logit (Best-Worst)—Vermunt and Magidson (2005)
2. MaxDiff model (Joint Best-Worst)—Marley and Louviere (2005)
3. MaxDiff Independence—Louviere (1993)

1. Best-Worst (sequential logit) model (used in this paper)

This approach models the best and worst alternatives as a sequential choice process (Bockenholt, 2002; Croon, 1989; Kamakura et al., 1994). That is, selection of the best option is equivalent to a first choice and then selection of the worst option is a (first) choice out of the remaining alternatives, where the worst choice probabilities are negatively related to the best utilities of these alternatives. This approach was used here, both in its standard form as well the extended SALC model form where the utilities are scale-adjusted to preference parameters. For more details, see Vermunt and Magidson, 2013.

2. MaxDiff Model (Marley and Louviere, 2005)

$$P_{ij}(s, t) = \frac{\exp(U_{ijs} - U_{ijt})}{\sum_{u=1}^K \sum_{v=1; v \neq u}^K \exp(U_{iju} - U_{ijv})}$$

This model is also known as the *joint best-worst* or *MaxDiff Quasi-Independence* model. It assumes best and worst options are selected simultaneously. That is, respondents are assumed to make their choices by evaluating all possible option pairs (Best, Worst), and selecting that pair having the largest difference in utilities (maximum difference). It is implemented in the Syntax module of LG Choice 5.0.²⁷

This model has a serious disadvantage from the others in that the *best* margin from the MaxDiff model is not consistent with the MNL model (see Marley and Louviere, 2005).²⁸

3. MaxDiff Independence Model

This model was proposed initially by Louviere (1993) and implemented by Sawtooth Software in its CBC/Latent Class program. Unlike the other two approaches that are based on true distributions, by relaxing the constraint that the most and least preferred options have to be different (Bacon et al., 2007), the probabilities from this model do not correspond to a correct theoretical probability distribution. As a result of this theoretical inconsistency, the BIC and related fit statistics for determining the number of classes are not appropriate for use with this

²⁷ This model can also be estimated with Sawtooth Software's CBC/HB package if the number of options is not too large (see Bacon et al. 2007).

²⁸ Both models 1 and 3 assume best and worst options are evaluated independently – not simultaneously. Both of these models yield best margins that are consistent with MNL.

model. Nevertheless, in practice this model generally yields parameter estimates that are similar to the sequential logit model (approach 1) when estimated with the same number of classes.²⁹

APPENDIX B: THE ONE-STEP VS. TANDEM APPROACH TO SEGMENTATION

In the widely used tandem approach, hierarchical Bayes (HB) analysis of MaxDiff responses is used in Step 1 to obtain individual utilities for each respondent, and in Step 2 these utilities are treated as continuous indicators in a LC Cluster model. By clustering the HB coefficients, one treats part of the individual variation as systematic (between clusters) and the remaining part as noise (within clusters).

In contrast, the standard 1-step LC choice model analyzes the choices from the MaxDiff exercise directly to determine the LC segments that differ with respect to their utilities. Again, the between-class heterogeneity is treated as systematic, and the within-class heterogeneity is treated as noise. Compared to the tandem approach, the 1-step approach with the same number of clusters picks up much more systematic variation, and with enough classes about the same amount as the HB model itself, so no systematic variation gets lost.

To demonstrate this result, we use the MaxDiff data provided by Lyon (2016), in his Case Study 1, who applied the tandem approach by obtaining HB utilities, and then estimating LC models having between 2 and 6 classes, settling on the 4-class model as best. The results from these LC models are presented in Table B1, with the standard deviation of the HB utilities as a measuring stick. The standard deviation statistic provides a useful measure of the total amount of heterogeneity that exists for each of the tasks.

As can be seen in column 1 of Table B1, (repeated in Tables B2, B3 and B4 where results are presented for various 1-step LC models), the HB utilities for six of the 14 tasks have relatively high standard deviations. Thus, we can say that these six tasks are responsible for most of the heterogeneity in the MaxDiff responses. We compare these HB standard deviations with the corresponding systematic standard deviations obtained under the tandem approach as well as under several 1-step LC approaches.

Table B1 compares these standard deviations (column 1) with the systematic (between-cluster) variability resulting from the clusters obtained from Step 2 of the 2-step HB/LC approach for models with 2–6 clusters. Note the following results:

- As the number of clusters increase, the standard deviations from the 2-step LC approach tend to increase but still remain far short of those of the HB utilities, even with 6 classes.
- The 4-class model (the one selected by Lyon) is fairly consistent with HB in identifying the tasks with the highest standard deviations (tasks A, D, F and J are highlighted in Table B1 compared to tasks B, D, F, H, J, and N identified by HB as most important).

²⁹ Allowing non-zero probabilities to be estimated for situation where the same alternative is selected as both best and worst violates the MaxDiff design that rules out this possibility (structural zeroes) and therefore always yield poor model fit according to the BIC. As a result, the BIC will always suggest more segments (classes) are needed, in an attempt to reduce these nonzero probability estimates to zero.

Table B1. Standard deviations for the HB individual utilities as reproduced in the second step of the 2-step approach.

Item	Standard Deviation					
	HB Observed	2-class	3-class	4-class	5-class	6-class
A	0.70	0.55	0.54	0.60	0.61	0.62
B	1.13	0.41	0.48	0.57	0.57	0.56
C	0.78	0.54	0.51	0.57	0.59	0.59
D	1.03	0.4	0.58	0.61	0.59	0.64
E	0.78	0.39	0.43	0.49	0.48	0.52
F	1.00	0.11	0.69	0.70	0.79	0.83
G	0.82	0.29	0.52	0.57	0.62	0.65
H	0.96	0.14	0.3	0.28	0.33	0.37
I	0.78	0.3	0.39	0.39	0.45	0.46
J	1.16	0.61	0.73	0.71	0.77	0.8
K	0.7	0.24	0.3	0.30	0.3	0.29
L	0.85	0.07	0.11	0.22	0.18	0.2
M	0.87	0.25	0.25	0.30	0.29	0.27
N	1.18	0.3	0.41	0.42	0.4	0.46
# parameters	119	43	58	73	88	103
BIC		20240	19715	19444	19264	19125

Table B2 below shows results obtained from the *standard* 1-step LC approach. Compared to Table B1:

- The standard deviation in Table B2 are higher and closer to those obtained from HB.
- The *highlighted* tasks are more consistent with those identified by HB. In particular, the 5 tasks selected as most important by the 4-class model (tasks B, D, F, J and N) correspond to 5 of the 6 tasks identified by HB as most important.

Thus, it is clear in this example that the 1-step approach yields segments that are more consistent with HB than the 2-step approach, a result consistent with other MaxDiff data that we have examined.

Table B3 shows 1-step LC results obtained when Scale-Adjusted LC (SALC) models are used instead of the standard LC models. The standard deviations are similar to those obtained by the standard LC model but the tasks with the highest standard deviations are even *more* similar to those obtained by HB.

Table B2. Standard deviations for standard latent class (1-step) best-worst.

Item	Standard Deviation					
	HB Observed	2-class	3-class	4-class	5-class	6-class
A	0.70	0.42	0.73	0.72	0.76	0.82
B	1.13	0.07	0.89	0.98	1.12	1.24
C	0.78	0.40	0.56	0.55	0.68	0.74
D	1.03	0.89	0.84	1.00	1.05	1.08
E	0.78	0.51	0.53	0.53	0.59	0.68
F	1.00	0.46	0.95	0.76	0.95	0.95
G	0.82	0.06	0.61	0.51	0.65	0.68
H	0.96	0.46	0.48	0.54	0.76	0.90
I	0.78	0.45	0.48	0.73	0.77	0.80
J	1.16	0.88	0.94	1.11	1.20	1.24
K	0.70	0.05	0.31	0.31	0.33	0.37
L	0.85	0.09	0.22	0.38	0.39	0.38
M	0.87	0.34	0.36	0.45	0.44	0.44
N	1.18	0.54	0.66	0.93	1.11	1.18
# parameters	119	27	41	55	69	83
BIC		25606	25047	24842	24635	24514

Table B3. Standard deviations for scale-adjusted latent class (SALC) best-worst.

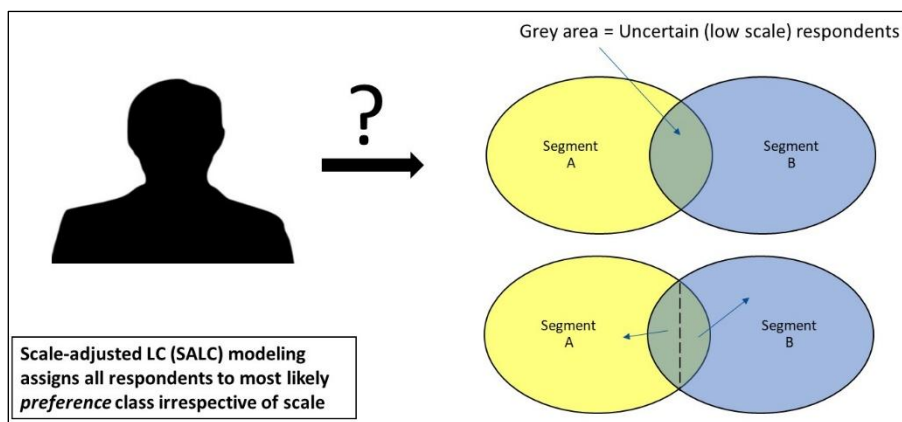
Item	Standard Deviation					
	HB Observed	2-class	3-class	4-class	5-class	6-class
A	0.70	0.44	0.62	0.61	0.74	0.82
B	1.13	0.13	1.04	1.01	1.27	1.36
C	0.78	0.38	0.44	0.47	0.63	0.75
D	1.03	0.87	0.98	0.94	1.03	1.09
E	0.78	0.50	0.48	0.49	0.58	0.68
F	1.00	0.52	0.92	0.84	0.82	0.87
G	0.82	0.08	0.54	0.50	0.54	0.59
H	0.96	0.47	0.42	0.53	0.67	0.84
I	0.78	0.49	0.27	0.74	0.74	0.79
J	1.16	0.93	0.73	1.17	1.19	1.26
K	0.70	0.04	0.21	0.20	0.34	0.34
L	0.85	0.04	0.33	0.37	0.34	0.38
M	0.87	0.30	0.51	0.45	0.48	0.56
N	1.18	0.65	1.08	1.01	1.06	1.23
# parameters	119	29	43	57	71	85
BIC		25283	24843	24645	24504	24387

APPENDIX C. THE SCALE-ADJUSTED LC (SALC) AND SALC TREE MODELS

With typical ratings data (e.g., Likert tasks) some respondents tend to rate all tasks relatively low while others tend to rate all tasks high. In a standard LC analysis of such data, the classes differ primarily on scale usage—one class consisting of persons with low ratings, while another consists of persons with high ratings. Since these classes are not useful for strategic purposes, a random intercept (Magidson and Vermunt, 2006; Popper et al., 2004) is included in the LC model to adjust for scale usage so that the resulting classes reflect meaningful differences in preference between tasks.

This type of scale usage problem does not occur with *choice-based tasks*. However, MaxDiff tasks, like other choice-based tasks, are subject to a different kind of scale problem. While not as serious as the one affecting ratings data, removing this scale confound requires a similar type of adjustment (Louviere and Eagle, 2006). The adjustment we use here is implemented in the Scale-Adjusted LC (SALC) model, introduced by Magidson and Vermunt (2007). For more details, see Groothuis-Oudshoorn et al. (2018).³⁰

Figure C1. SALC eliminates scale confound to obtain more meaningful segments.



This figure is a simplified illustration of how the SALC model eliminates the scale confound.

Segments A and B differ in their preferences. “Low scale” respondents may fall in a grey area where it is difficult to assign them into the most appropriate segment. SALC models recognize explicitly that respondents within any *preference class* differ with respect to scale and classify respondents into the most appropriate of these preference classes irrespective of scale.

Table C1 below is the SALC alternative to Table 8 presented in the body of this paper. Latent GOLD® was used to estimate both the LC and SALC models. (See Appendix A regarding LC models for MaxDiff data). Compared to the results from the standard 3-class model (Table 8), we see that the size of class 2 has now been reduced to 32% of respondents, re-assigning some of the low scale respondents into a class that is more consistent with their choices. Also, the objects (attributes) explaining the most heterogeneity—those with relatively high standard deviation (bolded in the Std. Dev. column)—are the same as in Table 8. However, the standard deviation of the class 2 preference parameters is now more comparable to those of the other classes.

³⁰ The scale problem was also recognized by Orme (2013). For additional applications of SALC models, see Burke et al. (2010).

Table C1. SALC model removes scale confound yielding preference parameters.

	Preference parameters			
Object name	Class 1	Class 2	Class 3	Std. Dev.
More frequent off-peak trains between major centers	0.9	-0.1	-0.1	0.52
Improved peak rail capacity	2.3	1.5	1.1	0.48
More frequent bus services on major routes	1.7	1.2	0.6	0.44
Extensions of light rail services	-0.8	-0.9	-1.2	0.17
Integrated fares	-0.4	-0.4	1.2	0.66
Integrated ticketing	-0.5	-0.6	1.5	0.84
Real-time arrival information	-0.7	-1.2	-0.4	0.31
New cycleways; more bike and scooter parking	-1.3	-0.8	-1.7	0.35
Trains use green power	-1.3	1.4	-0.9	1.21
Standard deviation	1.31	1.05	1.15	
Size	0.44	0.32	0.24	

The split of Class 1 into 2 child classes at level 2 of the LC Tree shows that these child classes maintain the theme of a preference for “More frequent off-peak trains,” but differ secondarily with respect to their preference for “Extensions of light rail services” (see Table C2).

Table C2. Preference parameters associated with the 2 child classes formed by splitting class 1.

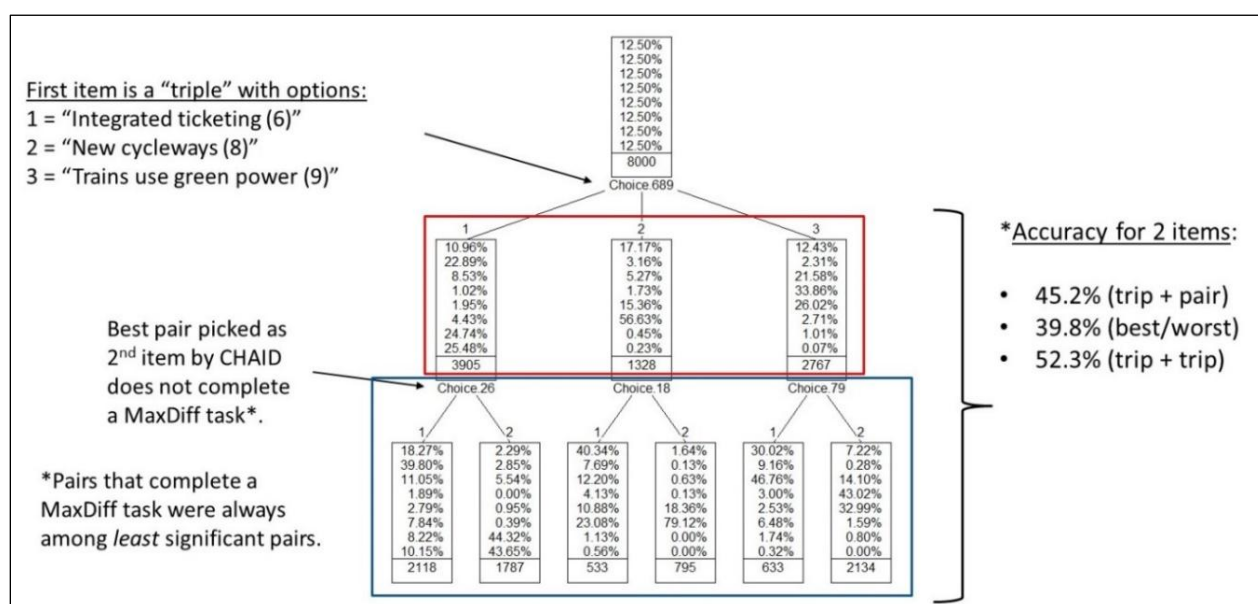
Object	Preference Parameters		Std. Dev.
	Class 11	Class 12	
More frequent off-peak trains between major centers	1.1	1.8	0.26
Improved peak rail capacity	4.0	2.2	0.63
More frequent bus services on major routes	2.5	2.5	0.03
Extensions of light rail services	-2.0	0.8	1.04
Integrated fares	-0.3	-0.8	0.19
Integrated ticketing	-0.5	-1.3	0.28
Real-time arrival information	-0.8	-1.5	0.27
New cycleways; more bike and scooter parking	-1.7	-2.2	0.19
Trains use green power	-2.2	-1.6	0.22
Standard deviation	1.45	1.27	
Size	0.28	0.16	

APPENDIX D: THE VALUE OF COLLECTING INFORMATION ON THE “LEAST IMPORTANT” RESPONSE

The value of collecting information on the “least important” response in MaxDiff exercises is well documented. To examine the value of collecting such information *in a typing tool* we performed a CHAID analysis as documented below.

In the case that MaxDiff tasks consists of triples, a MaxDiff choice task can be represented by selection of the *best* alternative from a triple followed by the best response to the relevant pair formed from the triple by eliminating the alternative selected as best. Since the pair consists of only 2 objects, the object *not* selected as best from the triple or pair would represent the least important among the three objects.

Figure D1. The accuracy for an adaptive typing tool using a sequential MaxDiff approach.



Using SI-CHAID in the exploratory mode, we split first on the triple {6, 8, 9} which is the most significant of all triples. Next, for each of the three resulting subgroups, we examined the p-value for all pairs, and split on the most significant of these pairs. The three pairs used to split were pair {2, 6}, pair {1, 6} and pair {7, 9}. None of these three pairs completed a (sequential) MaxDiff task. In fact, in all three cases, the pairs completing a MaxDiff task were ranked among the least significant of the pairs. The accuracy for the items is summarized in Figure D1.

Thus, we conclude that inclusion of MaxDiff tasks in a typing tool is not necessary to achieve high accuracy. Improved segment recovery (accuracy) in a typing tool can be expected by simply asking respondents to choose the best option from among two or more in each task.

REFERENCES

Bacon, L., Lenk, P., Seryakova, K., Vecchia, E. (2007). “Making MaxDiff More Informative: Statistical Data Fusion by Way of Latent Variable Modeling,” *Proceedings of the 2007 Sawtooth Software Conference*.

- Bockenholt, U. (2002). "A Thurstonian Analysis of Preference Change," *Journal of Mathematical Psychology*, 46, 300–314.
- Burke, P., Burton, C., Huybers, T., Islam, T., Louviere, J., and Wise, C. (2010). "The Scale Adjusted Latent Class Model: Application to Museum Visitation," *Tourism Analysis*, 15(2), 147–65.
- Croon, M. A. (1989). "Latent Class Models for the Analysis of Rankings," in G. De Soete, H. Feger, & K. C. Klauer (Eds.), *New Developments in Psychological Choice Modeling*, 99–121. North-Holland: Elsevier Science.
- Eagle, Thomas (2013). "Segmenting Choice and Non-Choice Data Simultaneously," *Proceedings of the 2013 Sawtooth Software Conference*.
- Groothuis-Oudshoorn, C.G.M., Flynn, T.N., Yoo, H.I., Magidson, J. and Oppe, M. (2018). "Key Issues and Potential Solutions for Understanding Health Care Preference Heterogeneity Free from Patient Level Scale Confounds," *The Patient: Patient-Centered Outcomes Research*, <https://rdcu.be/Mx8e>
- Kamakura, W.A., Wedel, M., and Agrawal, J. (1994). "Concomitant variable latent class models for the external analysis of choice data. *International Journal of Research in Marketing*," 11, 451–464.
- Komendant, L. (2016). Typing Tools in the Context of Choice Experiments, *Proceedings of the 2016 Sawtooth Software Conference*.
- Louviere, J.J. (1993). "The Best-Worst or Maximum Difference Measurement Model: Applications to Behavioral Research in Marketing," presentation at the American Marketing Association's 1993 Behavioral Research Conference, Phoenix.
- Louviere, J.J., and Eagle, T.C. (2006). "Confound It! That Pesky Little Scale Constant Messes Up," *Proceedings of the 2006 Sawtooth Software Conference*.
- Louviere, J.J., Flynn, T.N., and Marley, A.A.J. (2015). *Best-Worst Scaling: Theory, Methods, and Applications*, Cambridge: Cambridge University Press.
- Lyon, D. (2016). "Naïve Bayes Classifiers, or How to Classify via MaxDiff without Doing MaxDiff," *Proceedings of the 2016 Sawtooth Software Conference*.
- Magidson, J. (1994). "The CHAID Approach to Segmentation Modeling: CHi-squared Automatic Interaction Detection," in: Bagozzi, R. (ed.), *Advanced Methods of Marketing Research*. Blackwell.
- Magidson, J. (2003). Discussant comments on presentation by Steve Cohen, "Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation," *Proceedings of the 2003 Sawtooth Conference*.
- Magidson, J. (2018). "An improved latent class (LC) paradigm to obtain meaningful segments in the presence of scale confounds: Scale Adjusted Latent Class (SALC) Tree modeling," paper presented at the 2018 Advanced Research Techniques (ART) Forum.
- Magidson, J. and Bennett, G. (2016). "How to Develop a MaxDiff Typing Tool to Assign New Cases into Meaningful Segments," presentation at American Marketing Association's

- Advanced Research Techniques (ART) Forum, <https://www.statisticalinnovations.com/white-paper-maxdiff-typing-tool-final-2/>
- Magidson, J., Dumont, J., and Vermunt, J.K. (2015). "A New Modeling Tool for Identifying Meaningful Segments and Their Willingness to Pay: Improving Validity by Reducing the Confound between Scale and Preference Heterogeneity," presentation at American Marketing Association's Advanced Research Techniques (ART) Forum.
- Magidson, J., Eagle, T., and Vermunt, J.K. (2005). "Using Parsimonious Conjoint and Choice Models to Improve the Accuracy of Out-of-Sample Share Predictions," presentation at American Marketing Association's Advanced Research Techniques (ART) Forum.
- Magidson, J. (2018). "An improved latent class (LC) paradigm to obtain meaningful segments in the presence of scale confounds: Scale Adjusted Latent Class (SALC) Tree modeling," Proceedings of the 2018 Sawtooth Software Conference.
- Magidson, J., and Vermunt, J.K. (2006). "Use of Latent Class Regression Models with a Random Intercept to Remove Overall Response Level Effects in Rating Data," in A. Rizzi and M. Vichi (eds.), *Proceedings in Computational Statistics*, 351–360, Heidelberg: Springer.
- Magidson, J., and Vermunt, J.K. (2007). "Removing the Scale Factor Confound in Multinomial Logit Choice Models to Obtain Better Estimates of Preference," *Proceedings of the 2007 Sawtooth Software Conference*.
- Marley, A.A.J., and Louviere, J.J. (2005). "Some Probabilistic Models of Best, Worst, and Best-worst Choices," *Journal of Mathematical Psychology*, 49(6), 464–480.
- Orme, B. (2013). "Scale Constrained Latent Class," Research Paper Series, Sawtooth Software.
- Orme, B. and Johnson, R. (2009). "Typing Tools That Work," *Marketing Research*, Summer 2009.
- Popper, R., Kroll, J., and Magidson, J. (2004). "Applications of Latent Class Models to Food Product Development: A Case Study," *Proceedings of the 2004 Sawtooth Software Conference*.
- van den Bergh, M., Schmittmann, V.D., and Vermunt, J.K. (2017). "Building Latent Class Trees, with an Application to a Study of Social Capital," *Methodology*, 13(Supplement), 13–22.
- van den Bergh, M., van Kollenburg, G.H., and Vermunt, J.K. (in press). "Deciding on the Starting Number of Classes of a Latent Class Tree," *Sociological Methodology 2018*.
- Vermunt, J.K., and Magidson, J. (2005), *Technical Guide for Latent GOLD Choice 4.0: Basic and Advanced*, Belmont MA, Statistical Innovations.
- Vermunt, J.K., and Magidson, J. (2007). *Technical Guide for Latent GOLD 5.1.*, Belmont MA., Statistical Innovations.
- Vermunt, J.K., and Magidson, J. (2013). *Upgrade Manual for Latent GOLD Choice 5.0: Basic, Advanced, and Syntax*, Belmont MA: Statistical Innovations.
- Vermunt, J.K. (2013). "Categorical Response Data," in: M.A. Scott, J.S. Simonoff, and B.D. Marx (Eds.), *The SAGE Handbook of Multilevel Modeling*, Sage.

COMMENTS ON “DEVELOPMENT OF AN ADAPTIVE TYPING TOOL FROM MAXDIFF RESPONSE DATA”

THOMAS C. EAGLE

EAGLE ANALYTICS OF CALIFORNIA

The paper by Magidson and Madura continues the recent trend of developing typing tools for segmented choice data. Recent papers by Lyon (2016), Komendant (2016), and Orme and Johnson (2009) have shown various approaches to classifying new respondents into segments derived from choice data. The papers by Orme and Johnson and Lyon demonstrated their methods using MaxDiff models, which are a form of choice models. Komendant reviewed various methods using more traditional choice models. Magidson and Bennett also developed a typing tool using simulation methods for choice models as applied to MaxDiff (Magidson and Bennett, 2016). The methods varied with respect to issues such as:

1. What is the nature of the tasks a new respondent would see?
 - Are they tasks in the exact same form as those originally used in the modeling and segmentation? Or,
 - Are they a reduced form of the task? For example, fewer items in MaxDiff task (such as pairs), or fewer alternatives in a choice model.
2. How were the final set of tasks for the typing tool found?
 - Orme and Johnson (2009) use a Naïve Bayes classifier in the context of MaxDiff to create classification tasks of the original size. They use a greedy search algorithm to find the final set of tasks.
 - Lyon (2016) generalizes the above to allow for the construction of any size MaxDiff or choice modeling task. In addition, he discusses how to allow the addition of non-Maxdiff questions or respondent descriptors into the classifier.
 - Komendant (2016) reviews three tools: a simple pairwise classifier, a regression using rankings method, and the naïve Bayes classifier of Orme and Johnson coupled with a genetic algorithm to search for the best tasks with which to classify. The simple pairwise classifier did remarkably well when using a random forest method to classify respondents.
 - Magidson and Bennett (2016) use simulation methods and stepwise regression to find the best pairs of items to use to classify.

INNOVATIONS OF THE PRESENTATION

In their presentation, Magidson and Madura extend the work of Magidson and Bennett. They continue to use a simulation approach and make several innovative contributions. They replace the stepwise regression method to find the best items with which to classify new respondents with a decision tree approach. The decision tree approach produces an adaptive typing tool that yields fewer tasks per respondent, on average, and reduces the number of items each respondent might need to see in each task. The approach uses generated, simulation, data as the basis for developing the tree.

Further advancements include the innovative method of Latent Class Tree Modeling, which uses a CHAID-like method to develop choice model-based segments, and the use of scale adjusted latent classes to lessen the impact of scale in the construction of these segments. I find their new model, that combines the scale adjusted parameters and segment creation using the Latent Class Tree Modeling to be particularly compelling in the modeling of MaxDiff, and choice, data.

ISSUES THAT ARISE FROM THE PRESENTATION

How much do we gain by the effort required in using simulated data to generate the decision tree classifiers? Clearly, the Magidson and Madura approach shows the best results, but is all this effort really worth a few percentage point improvements? These improvements combined with those laid out by Komendant suggests the Magidson and Madura approach would be much better than the simple pairwise classifier even supplemented by decision tree methods such as CART, Random Forests, etc.

But, is the use of simulation enough to warrant the amount of work required to improve accuracy? The graphic in the Magidson and Madura paper, Figure 6, suggests a large improvement when moving from simulated adaptive paired comparisons to simulated adaptive triples (by 5-8% points), but an apple to apple comparison of the simulation approach using triples by the way of CART Trees or Random Forest trees has not been conducted. I suspect the improvement is meaningful, but further work should be done.

Another issue is: what is the best choice set of items/alternatives to show new respondents? Is it one that used generated tasks like those originally shown to respondents? Or, is it pairs? Or something else? This question arises from my history in using telephone interviews to classify new respondents. Using triples or more items per MaxDiff task were too much for respondents to handle. Pairs worked well and were easier to evaluate by respondents. Magidson and Madura clearly demonstrate that adding triples improves prediction a lot, but at the additional cost of more complex typing tasks.

Do we really need to ask both the best and worst in our MaxDiff classifiers? Magidson and Madura suggest we do not have to ask the worst choice. Using pairs eliminates this issue.

Performing a scale adjusted latent class tree model may yield segments that are easier to classify new respondents, but are the segments the best for management to use? This depends upon the way the segmentation results will be used. Magidson and Madura compare the results of a “standard” latent class choice model with that of a scale adjusted latent class segmentation. When clear-cut core “themes” emerge in the standard latent class choice model, then using a scale adjusted latent class tree model should result in more meaningful segments that are easier to classify new respondents into, and still capture the core “themes” found. But what if core “themes” do not emerge? What do we get in such a case? Are the splits in the scale adjusted latent class tree model meaningful in a managerially relevant way? I strongly believe in the use of scale adjusted MNL models, but there needs to be more applications of the latent class tree model to more data sets to evaluate whether such clear-cut themes always arise.

I do wish to re-emphasize a conclusion I made in 2013, and that the authors reiterate in their paper: if segmentation using Best-Worst data (or any choice data for that matter) is a primary objective of your project, DO NOT use the 2-step approach as described (See Eagle, 2013). One

should not first estimate hierarchical Bayesian utilities, rescale them, and then perform a segmentation analysis on these data. The process of running the hierarchical Bayes routine, which makes draws from the normally distributed upper level parameters, will not produce the same results as performing a latent class segmentation directly on the Best-Worst data (1-step approach).



Thomas C. Eagle

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*.
- Lyon, David (2016), "Naïve Bayes Classifiers, or How to Classify via MaxDiff without Doing MaxDiff," *Proceedings of the 2016 Sawtooth Software Conference*.
- Magidson, Jay and Gary Bennett (2016), "How to Develop a MaxDiff Typing Tool to Assign New Cases into Meaningful Segments," *Advanced Research Techniques Forum*, June, 2016.
- Orme, Bryan and Rich Johnson (2009), "Typing Tools That Work," *Marketing Research*, Summer 2009.

EXTENDING THE ENSEMBLE

CURTIS FRAZIER

ANA YANES

MICHAEL PATTERSON

RADIUS GLOBAL MARKET RESEARCH

ABSTRACT

Increasing variation in Step 1 clustering provides greater opportunity for efficiently discovering meaningful latent segments through ensemble analysis. This happens in step 1 by varying not only the algorithm and number of clusters, but also the variables that are input into the model itself. Finally, we compare the relative performance of this approach using synthetic data compared to existing approaches.

LIFE BEFORE ENSEMBLES

Developing segmentation solutions is perhaps the most fun and the most maddening portion of the marketing scientist's role. Segmentation is typically described as a mixture of art and science. This is because the beauty of a segmentation cannot be fully described by a single quantitative measure such as an R-squared or a hit rate.

However, the challenges on the "art" side are not our only challenges. On the science side, the analyst must make a myriad of decisions around inputs, techniques employed, and number of segments retained.

The current process for testing different combinations of these three elements to defining a segmentation is largely a matter of trial and error. Often, analysts will create various solutions using permutations of these three elements while simultaneously trying to understand not only the properties of the individual solution, but also how it relates to solutions using a different permutation.

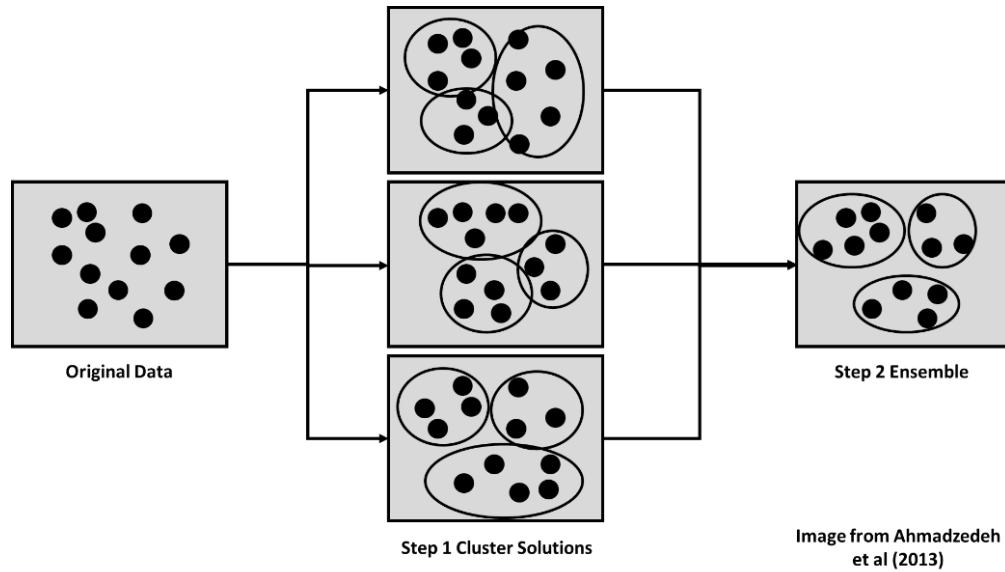
This approach is obviously an arduous and time-consuming proposition. In addition, the manual method provides many opportunities for both human error and simple failure to recognize patterns between solutions.

The reason we test multiple permutations is because we recognize that using a pre-determined set of inputs, techniques and number of segments is unlikely to result in stumbling across the optimal segmentation solution. Not only can different techniques produce different results, but the same technique can produce different results depending on the starting point (Fern and Brodley, 2004).

USING ENSEMBLES

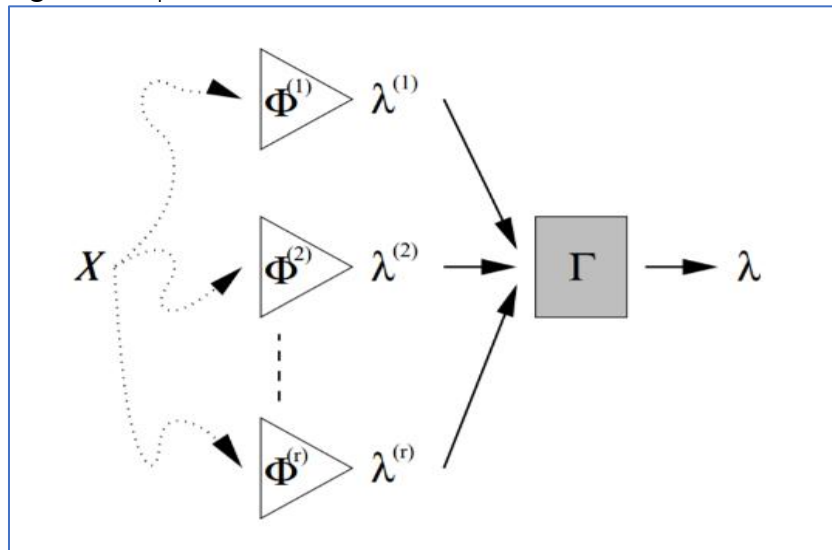
One solution to this trial-and-error approach is cluster ensembles. The idea behind ensembles is that while no individual solution may be optimal, we can combine solutions to understand the true structure of our latent segments.

In the image below, we can see how the same data can produce three separate solutions. With just three solutions and just two dimensions, it is easy to identify patterns in the data. By combining these three solutions based on their patterns of clustering, we can create an ensemble solution.



There are two methods of creating the ensemble solution:

1. Consensus method: Respondents are assigned to the segment to which they are clustered most often. This is the “four out of five dentists agree” approach. NOTE: because segments are nominal level metrics, in order for this method to work, we must re-order the solutions such that they align.
2. Clustering of the Clusters method: Rather than using a simple counting rule, this method inputs the multiple clustering solutions into a true multivariate model. This model can be based on any number of different algorithms (K-means, Latent Class, etc.). The figure below illustrates this process. Our individual level data (X) clusters into various solutions λ , using algorithms ϕ into consensus solution Γ .



The primary benefit of an ensemble solution is the mitigation of risks associated with a single solution. Because we don't have a particularly effective, objective measure of solution quality, it is difficult to know whether we have found an optimal solution using a single set of inputs, a single technique and/or a single number of solutions. By developing and combining multiple solutions, we can have more confidence that the segments exist. Fred and Jain (2002) refer to this as the process of evidence accumulation.

A secondary benefit of this process is that it removes extraneous variables, such as random starting points, from the equation. We want our solutions to be guided by attributes of interest, not by artifacts of the estimation process.

Finally, because ensemble solutions incorporate the finding of multiple models, it creates a more efficient process for the analyst. This should not be interpreted as a simple automation of the process, as ensemble solutions are not foolproof, nor will they necessarily find the optimal solutions.

ENSEMBLE IMPLEMENTATION

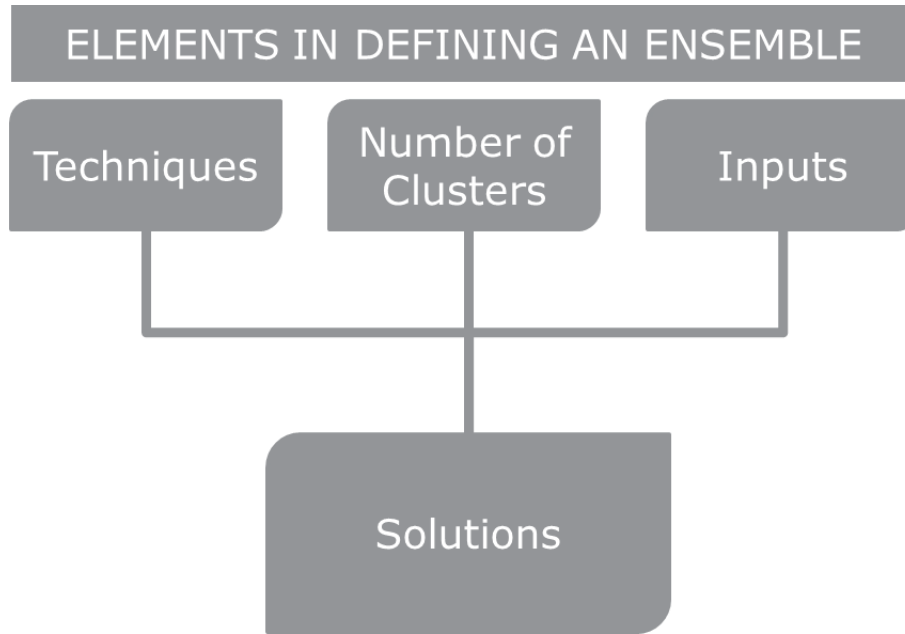
Sawtooth Software's Convergent Cluster Ensemble Analysis (CCEA) has become an important tool in allowing analysts to efficiently develop solutions manipulating both the techniques and the number of segments (Orme and Johnson, 2008). Similarly, the CLUE module in R (Hornik, 2005) provides an alternative method for the second step in generating an ensemble by relying on "evidence accumulation" (Fred & Jain, 2002).

Ahmadzede et al. (2013) provides a valuable review of various ensemble approaches and offers their own "graph-based approach for fuzzy partitions." (See also, Dietterich, 1999) Not only do each of these approaches provide a mechanism for evaluating and combining multiple cluster solutions, but the process of combining the solutions provides additional stability to the final selected solution (Lange et al., 2004).

Each ensemble approach is highly effective at the second step in generating ensembles—the clustering of the clusters. CCEA has the significant advantage of varying step 1 algorithms and numbers of segments, while processes such as CLUE focus exclusively on the ensemble development (Hornik 2005). However, in our experience, the biggest driver of differences in segmentation solutions (both quality and meaning) tend to be the inputs. In this paper, we will discuss a new approach in which we manipulate all three primary variables in a segmentation: the technique, the number of clusters and the inputs.

EXTENDED ENSEMBLES

Each of the existing techniques varies one or more of the three elements in developing solutions. But, none of them vary all three elements. For example, CCEA varies techniques and number of clusters, but keeps the inputs static while Strehl and Ghosh discuss varying number of clusters and inputs. Similarly, Fred and Jain (2002) discuss different ways of varying the elements, but settle on an approach in which the basic technique is kept constant, but different starting points and other technical parameters are allowed to vary.



Extended ensembles combine all three sets of inputs. Depending on the number of potential input combinations, we estimate randomly generated combinations of inputs using predetermined combinations of techniques and numbers of clusters.

The table below shows a portion of how this works. We see that Iterations 1–3 use the same set of inputs and number of segments but vary the techniques. Then, in Iterations 4–6 we keep those inputs static but begin to vary the number of segments. Iterations 6–9 begins the variance of the inputs. This process continues until we believe we have generated a sufficient number of candidate segmentation solutions for inclusion in the ensembling process.

Iteration	Technique	# Segments	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Attribute N
1	KM	4	✓		✓	✓		
2	LC	4	✓		✓	✓		
3	Hierarchical	4	✓		✓	✓		
4	KM	5	✓		✓	✓		
5	LC	5	✓		✓	✓		
6	Hierarchical	5	✓		✓	✓		
7	KM	4		✓		✓	✓	✓
8	LC	4		✓		✓	✓	✓
9	Hierarchical	4		✓		✓	✓	✓
...						
N						

The second stage of combining the candidate solutions into an ensemble can be done in various ways. For these tests, we used Latent Class clustering because of its flexibility in handling nominal level inputs without substantial data manipulation.

THE TESTS

Using synthetic data with known properties, we will show the degree to which an extended ensemble of solutions varying each of these variables can predict known segmentation membership. This will then be compared to alternative approaches, including traditional clustering, CCEA and Latent Class. Our approach mimics much of the CCEA approach but adds in the element of varying which “active” variables are included in the step 1 clustering.

Three synthetic datasets were created. These datasets each contained five known segments. Each segment had the same profile for each dataset. But, the size of each segment varied between the datasets.

SEGMENT DISTRIBUTIONS			
	Scenario 1: Even Distribution	Scenario 2: Slight Skew	Scenario 3: Extreme Skew
Segment 1	20%	30%	35%
Segment 2	20%	25%	30%
Segment 3	20%	20%	20%
Segment 4	20%	15%	10%
Segment 5	20%	10%	5%

Using each of these three datasets, we will determine the ability to recover the known segments using several methods:

1. Individual K-Means Clustering
2. Individual Hierarchical Clustering
3. Individual Latent Class Clustering
4. Ensemble K-Means Clustering
5. Ensemble Hierarchical Clustering
6. Ensemble Latent Class Clustering
7. CCEA
8. Extended Ensembles

For each clustering technique, we estimated a total of 500 iterations for 4- and 5-segment solutions for a total of 1000 K-means solutions, 1000 Hierarchical solutions and 1000 Latent Class solutions.

Our first test was to determine the ability for any single model estimation to recover our known segments. The table below illustrates the results. What is evident is that any individual solution is unlikely to recover the segments particularly well, with an average recover rate hovering around 35%.

Instead of looking at the average solution ability to recover the segments, we assume we stumble across the best case solution for each technique. In that scenario, each technique does a significantly better job at segment recovery. However, the likelihood of a particular technique finding that best combination of number of segments and inputs is small, particularly as our set of potential inputs increases.

		K-Means	Hierarchical	Latent Class
Even Distribution	Average	35.1%	35.1%	37.6%
	Best Case	53.1%	55.0%	52.8%
Slight Skew	Average	34.8%	33.2%	36.4%
	Best Case	52.7%	47.9%	52.2%
Extreme Skew	Average	34.7%	28.6%	35.8%
	Best Case	49.0%	44.5%	51.7%

Do we need to vary the techniques?

The second set of tests is to create ensembles within each technique. So, in these tests we are varying the number of clusters and the inputs but keeping the technique constant.

In this scenario, we see a substantial improvement in the ability of both K-Means and Latent Class to recover our known segments. This ability drops significantly for K-Means models as the distribution of segments becomes more skewed.

	K-Means	Hierarchical	Latent Class
	Ensemble	Ensemble	Ensemble
Even Distribution	67.8%	33.0%	76.5%
Slight Skew	66.2%	32.5%	76.3%
Extreme Skew	55.4%	27.1%	71.8%

Interestingly, Hierarchical models appear to perform no better when combined into an ensemble solution that they do individually. Somewhat counter-intuitively, this can happen when the solutions are either too similar to “accumulate evidence” or are so different that there isn’t enough relationships between the models to identify patterns.

The final test of existing techniques was to include CCEA as a modeling alternative. This was done in two ways.

1. The first test was an “all inputs” model where all potential input attributes were included in the model.
2. The second was a “variable inputs” in which a random subset of attributes were selected for 25 iterations. This “variable inputs” approach was tested to correspond to the approach used for each of the single technique tests.

	K-Means	Hierarchical	Latent Class	CCEA	CCEA
	Ensemble	Ensemble	Ensemble	All Inputs	Variable Inputs
Even Distribution	67.8%	33.0%	76.5%	77.1%	74.5%
Slight Skew	66.2%	32.5%	76.3%	74.8%	68.2%
Extreme Skew	55.4%	27.1%	71.8%	72.0%	65.4%

As we saw with the ensemble solutions with K-Means and Latent Class, both variations of CCEA are highly effective at recovering our known segments. The “all inputs” version is more effective than the “variable inputs” CCEA alternative. This is likely attributable to two aspects of the research design. First, our candidate set of solutions for the “variable input” models included just 25 randomly generated sets of input metrics. Second, our synthetic data included a total of 20 attributes, of which just 5 were designed to be irrelevant. Thus, an “all inputs” model will primarily include relevant variables, while in the real world our set of potential inputs may be heavily weighted towards less relevant metrics.

Our final segment recovery test is to include our Extended Ensembles approach. Rather than looking at varying inputs within a technique, we now create an ensemble across techniques. In this way, we are varying all three elements of the solution development.

The table below illustrates the results. What is evident is that creating ensembles across techniques (as with CCEA and Extended Ensembles) is generally more effective than ensembles within a technique. The exception to this finding is the strong performance of Latent Class ensembles, particularly when we have relatively equal segment sizes. Extended Ensembles appear to have the least drop-off in its ability to recover segments as segment size distributions become more skewed. However, overall, Extended Ensembles have not been found to be significantly better than CCEA models.

	K-Means	Hierarchical	Latent Class	CCEA	CCEA	Extended Ensembles
	Ensemble	Ensemble	Ensemble	All Inputs	Variable Inputs	
Even Distribution	67.8%	33.0%	76.5%	77.1%	74.5%	76.7%
Slight Skew	66.2%	32.5%	76.3%	74.8%	68.2%	74.2%
Extreme Skew	55.4%	27.1%	71.8%	72.0%	65.4%	73.4%

CONCLUSIONS

Combining multiple models into an ensemble, or “evidence accumulation,” appears to be highly effective at producing optimal solutions. This appears to be true whether we are looking within a single technique or varying the techniques. The key doesn’t appear to be what element(s) we are varying, but that we accumulate multiple models with which we can construct more information rich and informed segments.

Creating ensembles across techniques appears to be most important when our segment sizes are less evenly distributed. Obviously, in a real study, we do not know the number of segments, much less the distribution of segment sizes. For this reason, multi-technique ensembles like CCEA are a safer choice than single-technique ensembles.

Finally, while varying all three elements, as in Extended Ensembles, does not appear to be significantly better than CCEA except in cases with extreme skews in segment sizes, there is reason to believe that structure of our synthetic data may be inhibiting the ability for this approach to shine. Specifically, our synthetic data was structured to be heavily weighted towards relevant variables. In a standard 20-30 minute survey research project, we have a high likelihood of a greater proportion of irrelevant variables. In such a scenario, failure to vary the inputs should increase our likelihood of finding sub-optimal solutions.



Curtis Frazier



Ana Yanes



Michael Patterson

REFERENCES

- Ahmadzede, Mohammed, Zahra Azartash Golestan, Javad Vahidi and Babak Shirazi, “A Graph Based Approach for Clustering Ensemble of Fuzzy Partitions,” *Journal of Mathematics and Computer Science*, 2013.
- Dietterich, T.G. “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization.” *Machine learning*, volume 2, 2000.
- Fern, Xiaoli and Carla E. Brodley, “Cluster Ensembles for High Dimensional Clustering: An Empirical Study,” Oregon State University, 2006.
- Fred, A.L.N. and A.K. Jain. Data clustering using evidence accumulation. In *Proc. ICPR*, page to appear, 2002.
- Hornik, Kurt, “A CLUE for CLUster Ensembles,” *Journal of Statistical Software*, 2005.
- Kavšek, Branko, Nada Lavrač, and Anuška Ferligoj. Consensus decision trees: Using consensus hierarchical clustering for data relabelling and reduction. In *Proceedings of ECML 2001*, volume 2167 of *LNAI*, pages 251–262. Springer, 2001.
- Neumann, D.A. and V.T. Norton. Clustering and isolation in the consensus problem for partitions. *Journal of Classification*, 3:281–298, 1986a.
- Orme, Bryan and Rich Johnson, “Improving K-Means Cluster Analysis; Ensemble Analysis Instead of Highest Reproducibility Replicates,” *Sawtooth Software Research Paper Series*, 2008.

- Strehl, Alexander and Joydeep Ghosh, “Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions,” *Journal of Machine Learning Research*, Volume 3, Dec. 2002.
- Topchy, A.K. Jain, W. Punch, “Clustering Ensembles: Models of Consensus and Weak Partitions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 27, No 12, pp 1886–1881, Dec 2005.

SYNERGISTIC BANDIT CHOICE (SBC) DESIGN FOR CHOICE-BASED CONJOINT

BRYAN ORME
SAWTOOTH SOFTWARE

When studying concepts involving strong higher-order interaction effects, Synergistic Bandit Choice (SBC) for Choice-Based Conjoint can perform better than standard CBC studies. It is most useful for situations such as developing FMCG concepts involving aesthetic packaging (style, color, claims, packaging graphics, brand name, nutritional content, etc.) where it is expected that there may be strong and complex higher-order interaction effects (among 3+ attributes at a time) that are difficult to measure using traditional CBC design strategies. SBC leverages the collective knowledge of previously interviewed respondents, filters their choices to focus on the most significant interaction effects, and then oversamples the most synergistic feature combinations for evaluation by subsequent respondents.

Why refer to this method as a Bandit approach? The Wikipedia article on *Multi-Armed Bandit* (accessed 2/9/2016) states:

In probability theory, the multi-armed bandit problem is a problem in which a gambler at a row of slot machines (sometimes known as “one-armed bandits”) has to decide which machines to play, how many times to play each machine and in which order to play them. When played, each machine provides a random reward from a distribution specific to that machine. The objective of the gambler is to maximize the sum of rewards earned through a sequence of lever pulls.

The multi-armed bandit problem models an agent that simultaneously attempts to acquire new knowledge (called “exploration”) and optimize his or her decisions based on existing knowledge (called “exploitation”). The agent attempts to balance these competing tasks in order to maximize his or her total value over the period of time considered. There are many practical applications of the bandit model, for example portfolio design. The problem requires balancing reward maximization based on the knowledge already acquired with attempting new actions to further increase knowledge. This is known as the exploitation vs. exploration tradeoff in reinforcement learning.

For illustration of how the bandit problem relates to designing complex, synergistic products in marketing, let's consider a target FMCG (Fast Moving Consumer Goods) study for SBC with six attributes in a 24x20x6x8x12x4 design where there are 1.1 million possible product combinations. The concepts might be shown as integrated graphical elements (with transparent layers) allowing a concept to be displayed on-the-fly in a computerized survey environment by overlaying the graphics. Using the standard CBC approach, the observations would be especially

sparse if we wanted to examine the interaction effects. The interaction between the 24-level and 20-level attribute would involve 480 combinations *which standard CBC would sample evenly*. Even the interaction between the 12- and 8-level attributes involves 96 combinations. And this is without even considering higher-order interaction effects which could be meaningful in guiding an optimal FMCG package design with interactive and synergistic features. SBC takes a very different approach, using the key interaction effects (learned from prior respondents) as filters to identify and oversample synergistic joint features for new respondents. The highly preferred combinations across multiple attributes are more frequently investigated compared to the less preferred combinations, so we gain much higher precision than standard CBC to identify those complex level combinations that are near-optimal. SBC *exploits* prior information and *explores* new design combinations in a way that can lead to a much more efficient, targeted, and relevant choice survey experience than traditional CBC. The details are described in Appendix A.

SBC might involve showing each respondent around 8 to 20 CBC-looking tasks with somewhere between about 8 to 10 concepts per task, plus a None alternative. Standard discrete choice could be used; but we experimented with a multi-response format that may work even better when variety-seeking would be expected, such as for many food categories.

STRENGTHS OF THIS APPROACH

1. SBC focuses on discovering and exploiting synergistic interactions across multiple, typically categorical (nominal), product attributes. It discovers the synergistic combinations of product characteristics and focuses on alternatives with higher preference (first identifying key 2-way interaction effects, then later as more respondents complete the questionnaire identifying key 3-way interaction effects, etc.).
2. SBC can converge upon quite diverse products that serve different segments or occasions. For example, if half the people like granola power bars and the other half like chocolate bars, both desirable chocolate and granola bars will continue to be sampled and optimized, with supporting features that are specifically complementary to either chocolate or granola.
3. The use of a multi-check CBC task is more efficient for our purposes than discrete choice CBC as it collects more data and encourages within-respondent variety seeking for diverse optimal products that can satisfy different occasions and market segments.
4. The most preferred product features involved in significant interaction effects are soon identified and used as filters for discovering the most likely complementary supporting features. If two attributes are not important, then by definition their interaction Chi-square statistic will be small, since the observed and expected frequencies will both be flat, representing low signal.
5. SBC is extremely fast, well suited for on-the-fly computations during data collection.
6. Drawing new product combinations based on prior preferences and feature interactions leads to more efficient choice questionnaires that exploit the prior data, and as more data are collected, explores deeper (higher-order) interactions and complementary connections among a greater number of features.

PROOF OF CONCEPT PILOT TEST

To test the potential benefits of SBC, we recently conducted a pilot test of the SBC approach. We selected a product design situation where we would expect to see very highly interactive attributes. Such a design would violate the assumptions of standard additive, main effects + 1st order interactions models and could be difficult for the standard CBC design and estimation tools to deal with.

We asked respondents to evaluate various constructed phrases made up of levels from six different categorical (nominal) attributes. Here is the pilot test attribute list:

<u>Attribute 1:</u> 1 The young couple 2 The elderly couple 3 The children 4 The tourists 5 The young lovers 6 The teenagers	<u>Attribute 4:</u> 1 hills 2 beaches 3 forests 4 pastries 5 flowers 6 wines
<u>Attribute 2:</u> 1 climbed 2 strolled 3 gazed upon 4 smelled 5 tasted 6 enjoyed	<u>Attribute 5:</u> 1 of California 2 of France 3 of Oregon 4 of Hawaii 5 of Colorado 6 of Florida
<u>Attribute 3:</u> 1 the beautiful 2 the picturesque 3 the majestic 4 the satisfying 5 the delicious 6 the award-winning	<u>Attribute 6:</u> 1 in the summertime. 2 and took lots of pictures. 3 with their parents. 4 in the evening. 5 and posted the experience on Facebook. 6 and dreamed of their next adventure.

This attribute list may be used to construct $6^6 = 46,656$ total possible phrases (i.e., product concepts), many of which don't make any logical sense, such as:

The elderly couple tasted the delicious flowers of Oregon with their parents.

Some of the phrases make a great deal of sense, such as:

The young lovers tasted the award-winning wines of France in the evening.

or . . .

The children enjoyed the picturesque beaches of Hawaii with their parents.

Although it's pretty easy for any English-speaking human to see which combinations of these six attributes go well together without the need to conduct a market research study, such is not the case for designing consumer packaged goods with highly significant and complex interaction effects among categorical (nominal) attributes.

We asked our pilot respondents (Survey Sampling International Panelists) to complete 8 choice tasks, each with 10 concepts (plus a None). Rather than discrete choice, we employed a pick-any-3 approach:

Our computer has randomly arranged combinations of words to form phrases. We need a human like you to tell it which phrases make the most sense.

(Pick up to 3)	Which of these ten phrases make the most sense? (Pick up to 3)
<input type="checkbox"/>	The young couple strolled the majestic flowers of Hawaii and took lots of pictures.
<input type="checkbox"/>	The teenagers tasted the majestic pastries of Hawaii and posted the experience on Facebook.
<input type="checkbox"/>	The children enjoyed the beautiful flowers of France and took lots of pictures.
<input type="checkbox"/>	The tourists smelled the delicious wines of Colorado in the summertime.
<input type="checkbox"/>	The children tasted the award-winning hills of Oregon and posted the experience on Facebook.
<input type="checkbox"/>	The young lovers smelled the satisfying hills of Florida in the summertime.
<input type="checkbox"/>	The tourists gazed upon the delicious forests of Colorado with their parents.
<input type="checkbox"/>	The elderly couple strolled the picturesque wines of California in the evening.
<input type="checkbox"/>	The young couple climbed the award-winning beaches of Oregon and dreamed of their next adventure.
<input type="checkbox"/>	The young lovers gazed upon the satisfying forests of Florida with their parents.
<input type="checkbox"/>	None of these phrases make any sense

For a benchmark comparison, in addition to the choice tasks described above, below each pick-any-3 choice task respondents also completed a standard discrete choice CBC choice task where each respondent received a unique version (also known as a block), as shown below:

(Pick only ONE)	Which ONE of these ten phrases makes the most sense? (Pick just ONE)
<input type="radio"/>	The children smelled the beautiful wines of Hawaii in the summertime.
<input type="radio"/>	The elderly couple enjoyed the delicious beaches of France and took lots of pictures.
<input type="radio"/>	The teenagers gazed upon the picturesque flowers of Colorado in the summertime.
<input type="radio"/>	The young couple climbed the award-winning pastries of California and dreamed of their next adventure.
<input type="radio"/>	The tourists tasted the beautiful beaches of California and dreamed of their next adventure.
<input type="radio"/>	The young lovers strolled the majestic flowers of Colorado with their parents.
<input type="radio"/>	The young lovers strolled the majestic wines of Florida with their parents.
<input type="radio"/>	The teenagers gazed upon the picturesque hills of Florida and posted the experience on Facebook.
<input type="radio"/>	The tourists tasted the satisfying forests of Oregon in the evening.
<input type="radio"/>	The young couple smelled the satisfying hills of Hawaii and posted the experience on Facebook.
<input type="radio"/>	None of these phrases make any sense

For this pilot test, we didn't undertake the considerable effort to automate the SBC adaptive design algorithm and integrate it within our survey software platform, so we conducted the pilot test in small data collection batches and manually performed the design algorithm (using macros in Excel) to generate the design for the next batch of respondents.

We paused the data collection after collecting the first 31 respondents, discarded two of the respondents who were obviously speeding, and examined each of the 15 possible two-way

interaction tables among the 6 attributes. The most significant interaction according to the Chi-Square statistic¹ was between attributes 2 and 4, with raw frequencies of choice as follows:

Wave 1 Counts (Most Significant Interaction Effect) n=29

	hills	beaches	forests	pastries	flowers	wines
climbed	26	3	11	0	0	0
strolled	15	25	15	0	8	2
gazed upon	32	27	26	3	27	5
smelled	3	1	5	6	28	3
tasted	0	0	1	20	0	10
enjoyed	31	31	19	20	28	15

For the most part, this table of frequencies looks very reasonable, except for a few instances (such as one respondent who seemed to think the idea of *tasting the forests* made sense or the two instances in which *strolling the wines* was chosen). These are complex choice tasks and respondents answer with error, so this is to be expected. The most frequently chosen combination (chosen 32 times) was the couplet *gazed upon the hills*.

With likelihood proportional to the frequencies above (after adding 5 to each cell of the table), we drew a few thousand pairs of levels from attributes 2 and 4 to show the next batch of about 30 respondents. For each pair, we counted the observed choices containing that level couplet and drew levels for the remaining four attributes with likelihood proportional to their filtered marginal counts² (after adding 5 to each raw count). For example, let's imagine a draw of the pair *climbed the hills*. Just isolating the 26 times respondents selected *climbed the hills*, we counted how many times the other six levels of each of the other four attributes (attributes 1, 3, 5, and 6) were chosen, then drew levels (with likelihood proportional to the observed frequencies) to combine with the couplet *climbed the hills* to form new sentences (i.e., product concepts). To those thousands of drawn sentences, we added a few hundred randomly drawn additional concepts such that 20% of the designed concepts for the second batch of respondents would be randomly constructed. These random concepts serve two purposes. First, they limit the possibility that successive iterations might move in the wrong direction because of unrepresentative early samples. Second, by retaining a number of items which make little sense, respondents in later waves are not confronted with a list where all of the items make sense. In that case the task might become overly difficult and less rewarding for a respondent.

Recall that respondents were asked to complete eight standard CBC tasks as well (placed on the same screen under the eight SBC tasks). For the next wave of respondents, we also regenerated the standard CBC design plan using a different random seed such that the second and later batches of respondents would not simply get a repeat of the first batch of respondents' choice tasks.

After the second wave of data collection (and after cleaning any obvious speeders) we had a total of 58 respondents. The interaction between attributes 2 and 4 was again the strongest interaction effect. The frequencies for that interaction table were:

¹ We computed the Chi-Square statistic after adding a raw frequency of 5 counts to each cell of the interaction tables.

² Although it seems like we are just using 2-way interactions at this stage to draw new product concepts for the next respondents to evaluate, we are actually using 3-way interactions. The 2-way couplets are used as filters to observe how frequently that couplet was chosen with levels of each other attribute. This involves examining 3-way frequencies to draw new concepts with likelihoods according to those observed 3-way effects.

Wave 2 Counts (Most Significant Interaction Effect) n=58

	hills	beaches	forests	pastries	flowers	wines
Climbed	62	8	22	1	0	1
Strolled	37	73	36	7	19	3
gazed upon	76	62	57	9	60	9
Smelled	4	4	10	8	61	7
Tasted	3	4	3	45	1	31
Enjoyed	71	89	46	35	64	32

We hadn't yet reached our frequency threshold of 100 for any one cell within this interaction table (the max count is 89), so we retained this most significant interaction table for drawing the product concepts for the next 30 respondents (rather than graduating to a 3-way interaction table).

After the third wave of data collection (and after cleaning any obvious speeders) we had a total of 87 respondents. We now had enough data to eclipse the 100 count frequency target for a cell within the most significant 2-way interaction, so we felt we had enough data to graduate to the most significant 3-way interaction effect (among attributes 2 x 3 x 4). The frequencies for that 3-way most significant interaction table were:

Wave 3 Counts (Most Significant Interaction Effect) n=87

	hills	beaches	forests	pastries	flowers	Wines
climbed the beautiful	14	1	3	0	0	1
climbed the picturesque	22	3	7	0	0	0
climbed the majestic	31	1	7	0	0	0
climbed the satisfying	6	4	3	0	0	1
climbed the delicious	3	0	3	0	0	0
climbed the award-winning	16	1	4	1	0	0
strolled the beautiful	19	40	8	0	3	0
strolled the picturesque	8	34	16	2	3	1
strolled the majestic	16	20	13	2	6	1
strolled the satisfying	7	15	4	1	2	0
strolled the delicious	3	4	1	2	0	1
strolled the award-winning	8	17	8	0	7	0
gazed upon the beautiful	25	21	21	3	37	3
gazed upon the picturesque	36	19	15	0	16	0
gazed upon the majestic	34	35	35	1	14	2
gazed upon the satisfying	7	11	9	1	9	2
gazed upon the delicious	5	0	0	1	2	1
gazed upon the award-winning	13	8	10	4	24	1
smelled the beautiful	2	0	3	0	20	0
smelled the picturesque	0	1	1	0	11	1
smelled the majestic	2	0	3	0	24	1
smelled the satisfying	0	0	1	4	18	1
smelled the delicious	1	1	0	3	2	4
smelled the award-winning	0	2	3	2	22	3
tasted the beautiful	1	0	2	4	0	7
tasted the picturesque	1	0	0	3	0	3
tasted the majestic	1	1	0	8	1	5
tasted the satisfying	0	1	0	11	1	7
tasted the delicious	0	0	0	17	0	17
tasted the award-winning	1	2	1	23	0	7
enjoyed the beautiful	36	38	16	7	35	5
enjoyed the picturesque	43	44	25	1	34	1
enjoyed the majestic	19	41	18	4	9	7
enjoyed the satisfying	9	14	5	1	9	8
enjoyed the delicious	1	4	2	15	3	19
enjoyed the award-winning	7	20	6	12	16	10

At this point, the most commonly picked 3-way attribute combinations for the most significant 3-way interaction effect are *enjoyed the picturesque beaches* (44 count), *enjoyed the picturesque hills* (43 count), and *enjoyed the majestic beaches* (41 count). But, there are other combinations of attributes that also are quite complementary and seem to make a lot of sense to the respondents. Plus, we still need to figure out what other three features from attributes 1, 5, and 6 go best with these most preferred combinations. On to wave 4, this time using the 3-way interaction table above as a filter to make draws³ for the most likely combinations of attributes 1, 5, and 6.

³ Although it seems like we are using 3-way interactions at this stage to draw new product concepts for the next respondents to evaluate, we are actually using 4-way interactions. The 3-way triplets are used as filters to observe how frequently that triplet was chosen with levels of each other attribute. This involves examining 4-way frequencies to draw new concepts with likelihoods according to those observed 4-way effects.

After wave 4 (n=115), the frequencies for the 3-way most significant interaction table were:

Wave 4 Counts (Most Significant Interaction Effect) n=115

	hills	beaches	forests	pastries	flowers	wines
climbed the beautiful	16	1	4	1	0	1
climbed the picturesque	33	3	7	0	0	0
climbed the majestic	40	2	8	0	0	0
climbed the satisfying	7	5	3	0	0	2
climbed the delicious	3	0	3	0	0	0
climbed the award-winning	25	1	7	1	0	0
strolled the beautiful	25	58	12	0	3	0
strolled the picturesque	10	52	23	3	4	1
strolled the majestic	22	23	18	2	7	1
strolled the satisfying	8	19	5	2	2	0
strolled the delicious	4	4	1	3	0	1
strolled the award-winning	12	22	10	0	7	0
gazed upon the beautiful	35	27	30	3	56	4
gazed upon the picturesque	46	32	18	0	26	0
gazed upon the majestic	47	44	44	1	16	2
gazed upon the satisfying	11	15	12	1	14	3
gazed upon the delicious	5	0	1	2	2	1
gazed upon the award-winning	14	10	12	4	34	2
smelled the beautiful	2	1	6	0	25	0
smelled the picturesque	1	1	2	0	17	1
smelled the majestic	2	0	3	0	32	1
smelled the satisfying	1	0	2	4	20	1
smelled the delicious	1	1	1	4	3	4
smelled the award-winning	0	4	3	2	24	4
tasted the beautiful	1	0	2	4	0	14
tasted the picturesque	2	0	0	5	0	4
tasted the majestic	1	2	0	10	1	7
tasted the satisfying	0	2	0	14	1	9
tasted the delicious	0	0	0	23	0	23
tasted the award-winning	1	2	1	28	0	9
enjoyed the beautiful	51	53	22	7	47	6
enjoyed the picturesque	58	65	35	1	38	1
enjoyed the majestic	26	53	24	4	10	8
enjoyed the satisfying	9	19	5	2	9	11
enjoyed the delicious	1	5	2	23	3	23
enjoyed the award-winning	11	27	7	15	21	16

After wave 5 (n=143), the frequencies for the 3-way most significant interaction table were:

Wave 5 Counts (Most Significant Interaction Effect) n=143

	hills	beaches	forests	pastries	flowers	Wines
climbed the beautiful	24	1	4	1	0	1
climbed the picturesque	39	3	7	0	0	0
climbed the majestic	50	2	9	0	0	0
climbed the satisfying	8	6	3	0	0	3
climbed the delicious	3	0	3	0	0	0
climbed the award-winning	30	1	8	2	0	0
strolled the beautiful	30	78	14	0	3	0
strolled the picturesque	16	70	29	3	4	2
strolled the majestic	26	30	21	2	8	1
strolled the satisfying	9	26	5	2	2	0
strolled the delicious	4	5	1	3	0	1
strolled the award-winning	14	26	12	0	8	0
gazed upon the beautiful	44	36	37	3	76	4
gazed upon the picturesque	61	44	20	0	32	1
gazed upon the majestic	65	53	49	1	20	4
gazed upon the satisfying	12	15	13	2	17	3
gazed upon the delicious	7	0	1	2	4	1
gazed upon the award-winning	14	10	14	4	40	2
smelled the beautiful	3	2	6	0	28	0
smelled the picturesque	1	1	3	1	24	1
smelled the majestic	2	0	3	0	37	1
smelled the satisfying	2	0	3	4	23	2
smelled the delicious	2	1	1	5	5	4
smelled the award-winning	0	4	4	3	31	4
tasted the beautiful	1	0	3	5	0	20
tasted the picturesque	2	0	0	5	0	4
tasted the majestic	1	2	0	11	1	8
tasted the satisfying	0	2	0	16	1	12
tasted the delicious	0	0	0	28	0	25
tasted the award-winning	1	2	1	34	0	11
enjoyed the beautiful	57	73	29	8	58	6
enjoyed the picturesque	74	87	40	1	42	1
enjoyed the majestic	30	64	26	4	11	11
enjoyed the satisfying	10	19	5	2	11	12
enjoyed the delicious	1	7	2	29	3	28
enjoyed the award-winning	13	40	8	17	26	17

Now that we've collected the data, there are different ways to build models and predict the best concept combinations. These are the approaches we used:

Data Source:	Modeling Approach:
Synergistic (Bandit) CBC	<ul style="list-style-type: none"> - Aggregate Logit with all significant 2-way Interaction Effects - Counting-Based Simulator (using the same logic as the design generation steps)
Traditional CBC	<ul style="list-style-type: none"> - Aggregate Logit with all significant 2-way Interaction Effects - HB with Top Two 2-way Interaction Effects

For the aggregate logit approach, we identified multiple 2-way interaction effects⁴ to add to the main-effects specification based on a forward stepwise 2-Log Likelihood approach, with 95% confidence level threshold for inclusion. Three-way interaction effects were not possible to estimate given the limitations of the data and the number of parameters involved in a 3-way interaction term, though we could have remedied that for a larger real study by collapsing interactive attributes into a single factor and recoding any levels exhibiting zero or extremely low frequency of choice into a single catch-all level of that collapsed factor (as described below in footnote #4). The best internal fit model specification including all significant 2-way interaction terms was Main Effects: Att1, Att2, Att3, Att4, Att5, Att6; plus Interaction Terms: Att2xAtt4, Att3xAtt4, Att4xAtt5, Att1xAtt6, Att2xAtt3, Att1xAtt4. As you can see, there were many interaction effects that passed the 95% confidence threshold and continued to improve the LL fit. This led to 30 main-effects parameters + 150 interaction parameters + None for a total of 181 parameters in the aggregate logit model.

Note that for the Synergistic (Bandit) CBC, we ran aggregate MNL under the assumption of chip-allocation (each respondent had 3 chips to allocate among the 10 concepts plus the None; if a respondent picked just 1 non-None concept, then 2 chips were assumed to be given to the None, etc.). For the Traditional CBC data, we ran the standard aggregate MNL.

The reader may naturally wonder how well suited the SBC data were for traditional MNL modeling. Despite using such an aggressive adaptive design strategy that oversampled the most synergistic level combinations, the overall design efficiency (considering all attribute levels) in terms of main effects for SBC was 65% as efficient as traditional level-balanced, orthogonal CBC. (SBC oversamples most preferred levels, so the efficiency for specific most-preferred levels should be enhanced relative to CBC.) The correlation in aggregate logit main effects between SBC and CBC was 0.93, which is high given our limited sample size and suggests that any selection bias for SBC is quite minimal.

For the HB modeling, we believed that a 150-parameter model would be too demanding given the sparse data conditions of n=143 respondents. Rather, we added the two most important interactions to the model (Att2xAtt4; Att3xAtt4), for a total of 30 main effects parameters + 50 interaction terms + None = 81 parameters in the model.

To assess face validity, we enumerated the 100 top concepts found (separately) by the SBC and CBC methods (it's simple and extremely quick to conduct an exhaustive simulation search

⁴ As a technical note, aggregate MNL will not converge if any 2-way combination of levels was picked either 100% or 0% of the time (the latter is the more likely outcome with these models). To resolve this, you could simply append a few synthetic tasks to the bottom of the data set wherein a synthetic respondent "saw" each combination of the attributes within the same task and placed an allocation of 1 on each concept and 0 for the None. Another approach is to collapse attributes involved in interactions into single combinatorial factors and then to recode any extremely tiny or 0 probability levels into a single catch-all generic level, which would save many parameters to estimate in the model.

across 46,656 possibilities⁵). We then counted how many of these top 100 phrases included any illogical attribute combinations. For standard CBC with the aggregate logit solution (with extensive 2-way interaction effects), if you believe that elderly people posting to Facebook in 2016 (when the data collection occurred) is illogical, then 6 of the 100 top combinations lacked face validity. Otherwise, all 100 were perfectly logical. For the SBC aggregate logit solution (with extensive 2-way interaction effects) 11 of the top 100 combinations involved elderly people posting to Facebook. Otherwise, all 100 were perfectly logical. To us, this seemed like very good face validity for both methods. No optimal combination had phrases involving such things as *tasting flowers*, *smelling beaches*, or *climbing pastries*. While this is a satisfying outcome, we wanted a more quantitative assessment of validity.

To test the success of the four different approaches (2 data collection methods x 2 models), we imagined a situation in which the researcher wants to design the best total phrase (product concept) that goes together with each of the six levels of Attribute 4:

- Hills
- Beaches
- Forests
- Pastries
- Flowers
- Wines

Using each approach we exhaustively searched across all 46,656 product combinations (takes about 10 seconds using Sawtooth Software Advanced Simulation Module) to find the best concepts containing each level of attribute 4. There was no agreement among the four approaches regarding which product concepts were the *most* ideal to accompany these six levels of attribute 4. So, we took the “optimal” concepts found by each of the four approaches and specified them as composite items (phrases) within a MaxDiff design to field among an additional and new wave of about 50 respondents. So that respondents didn’t always face such a hard MaxDiff task (since all these are probably quite good concepts), we added six purely random items to the design, for a total of 30 items in the MaxDiff experiment. Each respondent completed 20 MaxDiff tasks, where each task showed 3 items at a time.

For example:

Our computer has randomly arranged combinations of words to form phrases. We need a human like you to tell it which phrases make the most sense.

Which of these 3 phrases makes the Most Sense and which makes the Least Sense?

(1 of 20)

Most Sense	Least Sense	
<input type="radio"/>	<input type="radio"/>	The tourists enjoyed the delicious wines of Oregon in the summertime.
<input type="radio"/>	<input type="radio"/>	The children enjoyed the delicious hills of Florida and took lots of pictures.
<input type="radio"/>	<input type="radio"/>	The young couple enjoyed the picturesque hills of Colorado and dreamed of their next adventure.

⁵ For search spaces larger than about 50MM potential outcomes, a heuristic search could be useful instead of an exhaustive search. But, exhaustive could work quite well for up to at least 50MM potential product combinations.

The table below shows the MaxDiff aggregate logit scores (rescaled to probability scaling) for the 30 items.

Results of MaxDiff Validity Test (n=53)

Concepts Identified via Synergistic CBC, Aggregate Logit with all Significant 2-way Interactions:

The young couple enjoyed the picturesque hills of Colorado and dreamed of their next adventure.	4.2
The tourists enjoyed the beautiful beaches of Hawaii and posted the experience on Facebook.	6.3
The tourists gazed upon the majestic forests of Oregon and posted the experience on Facebook.	5.0
The teenagers tasted the award-winning pastries of France with their parents.	3.3
The young couple gazed upon the beautiful flowers of Hawaii in the summertime.	3.5
The young couple enjoyed the delicious wines of California and dreamed of their next adventure.	5.0

Geometric Mean: 4.44

Concepts Identified via Synergistic CBC, Counting Analysis Approach:

The teenagers enjoyed the picturesque hills of France in the evening.	2.3
The elderly couple enjoyed the picturesque beaches of Hawaii and took lots of pictures.	6.2
The children gazed upon the majestic forests of Oregon and took lots of pictures.	3.1
The young couple tasted the award-winning pastries of France and posted the experience on Facebook.	5.0
The teenagers gazed upon the beautiful flowers of Florida in the summertime.	3.5
The tourists enjoyed the delicious wines of Oregon in the summertime.	3.3

Geometric Mean: 3.68

Concepts Identified via Standard CBC, Aggregate Logit with all Significant 2-way Interactions:

The young couple enjoyed the beautiful hills of Colorado and posted the experience on Facebook.	5.2
The elderly couple strolled the beautiful beaches of Florida and took lots of pictures.	5.3
The young lovers enjoyed the beautiful forests of Colorado and dreamed of their next adventure.	3.4
The elderly couple tasted the satisfying pastries of France and took lots of pictures.	2.0
The elderly couple enjoyed the beautiful flowers of Hawaii and took lots of pictures.	5.7
The elderly couple tasted the delicious wines of California and took lots of pictures.	3.1

Geometric Mean: 3.85

Concepts Identified via Standard CBC, HB with Top Two Significant 2-way Interactions:

The young couple strolled the beautiful hills of California and took lots of pictures.	4.7
The young couple strolled the beautiful beaches of California and took lots of pictures.	6.1
The young couple gazed upon the beautiful forests of California and took lots of pictures.	3.9
The young couple enjoyed the satisfying pastries of California and took lots of pictures.	1.2
The young couple gazed upon the beautiful flowers of California and took lots of pictures.	3.7
The young couple tasted the delicious wines of California and took lots of pictures.	3.4

Geometric Mean: 3.45

Randomly Generated Concepts:

The children enjoyed the delicious hills of Florida and took lots of pictures.	0.3
The tourists smelled the picturesque beaches of Colorado in the summertime.	0.2
The children strolled the satisfying forests of Florida with their parents.	0.6
The young couple strolled the delicious pastries of California and dreamed of their next adventure.	0.2
The young couple climbed the award-winning flowers of Florida with their parents.	0.1
The young couple strolled the award-winning wines of Florida and dreamed of their next adventure.	0.4

Geometric mean: 0.24

Of the four approaches, the SBC approach with aggregate logit modeling found the most optimal phrase, though it only barely edged out a couple items found by other approaches. On average⁶, the SBC approach did better (4.44 vs. 3.85; $p=0.014$)⁷ than the traditional CBC approach (though we must not lose sight of the fact that SBC had an advantage in that it collected more data per respondent due to the multi-select format).

⁶ Since the scores are on a probability scale, taking the geometric average is more appropriate than taking the arithmetic mean and punishes an approach more for obtaining a particularly low probability score on any one item.

⁷ Separately, we used HB to compute the scores so we could perform a Bayesian test of significance (by counting the alpha draws). This test showed with 98.6% confidence ($p=0.014$) that the SBC Logit approach led to the identification of near-optimal items (product concepts) that were preferred to the items found by the CBC Logit approach.

For both types of data collection (SBC and CBC) the aggregate logit with all significant 2-way interaction effects tended to work better than the HB approach—though we limited the HB approach to using only the top 2 interaction effects for fear of overfitting. We should note that respondents were probably much more homogeneous regarding their evaluation of this attribute list than respondents might be for typical market research product categories, which was an advantage for the aggregate logit approaches. For the SBC data, the logit model worked better than the counting-based simulator. As a pilot test, this is only an initial proof of concept using a relatively small sample size.

CONCLUSIONS

We embarked on this research believing that traditional CBC would not do a very good job identifying the combinations of words that made up ideal phrases where there were so many high-order interaction effects. Though we cannot know for certain what the optimal product concepts actually were, it appears to us that traditional CBC with the aggregate logit approach with all significant 2-way interactions specified worked reasonably well for this smallish-scope pilot study. But the SBC approach did even better ($p=0.014$). With larger sample sizes, collapsing attributes, and recoding low-frequency combinations into catch-all categories, 3-way or potentially even higher-order interactions could be included as well. When so many interactions are in play, HB individual-level estimation probably has less to offer due to the real possibility of overfitting. A latent class approach could be a good compromise.

The SBC approach makes some key departures from the standard CBC approach to tackle the problem of highly synergistic products:

- Encouraging some attributes with relatively many levels (say, 10 to 30 levels) to investigate a very large array of potentially beneficial product features (though the researcher should still do everything possible to be judicious and limit the attribute levels in the study based on prior information; each additional level added to the design costs something).
- Welcoming as many prohibitions between attributes that the researcher for certain can rule out *a priori*. Because the emphasis of the analysis is on modeling interactive combinations of attributes, this can significantly reduce the design space (attributes involving interactions must be collapsed into single factors prior to running MNL).
- Abandoning level-balanced, orthogonal plans in favor of an adaptive bandit procedure that aggressively oversamples combinations of attributes that are synergistic and important.
- The use of a multi-check CBC questionnaire format to collect more data and accommodate variety-seeking for products satisfying multiple occasions. Even though our pilot study was not designed to be able to isolate and test the accuracy of the multi-check versus a discrete choice format, our feeling is that the multi-check format should perform better in these kinds of design applications.
- During the MNL modeling, encouraging as many 2-way interaction terms (and potentially 3-way interaction terms) that are statistically significant. Even though the number of parameters to estimate have increased geometrically compared to standard main-effects CBC, the relatively few parameters that matter most in explaining

respondents' highest multi-feature preferences are vastly oversampled compared to the non-significant ones. This means that even though the average number of observations relative to parameters estimated is much lower for these models than traditional CBC, the concentration of observations on the relatively few parameters that actually make up the near-optimal products is many, many times greater than with standard CBC.

- Due the number of parameters included in these MNL models (often in the 100s due to the interaction effects), emphasizing the use of aggregate logit or latent class instead of HB.

SBC creates an experimental design that aggressively oversamples the most synergistic and preferred combinations of features. On the flip side, this means that SBC would be a bad approach when the goal is to identify and discriminate among the worst combinations of features. When the subject of the study involves highly interactive collections of product features, our belief is that the adaptive bandit design approach coupled with the emphasis on specifying very large models covering potentially 100s of interaction parameters should lead to more precise identification of near-optimal concepts than traditional CBC. A presentation at the 2015 Sawtooth Software Conference regarding Bandit MaxDiff suggests this outcome (Fairchild, Orme, and Schwartz, "Bandit Adaptive MaxDiff Designs for Huge Number of Items," 2015 Sawtooth Software Conference). How much better and whether it is worth the effort of doing the bandit design strategy for CBC remains to be proven with a larger study (though, this pilot test is encouraging).

FUTURE RESEARCH

Our reviewer, Joel Huber, suggested that rather than select the one most important interaction effect and use it as the filter by which a new product is drawn, we could draw new couplets (triplets, etc.) from all possible two-way (three-way, etc.) interactions probabilistically, with likelihood proportional to $(1-p)f$, where p is the p -value from the interaction effect Chi-square test and f is the relative frequency within the selected two-way (three-way, etc.) table. Further work needs to be done regarding how to compute the Chi-Square statistic when prohibitions between attributes are involved. And, of course, validation using a larger sample is called for as the work presented here involved a limited pilot study.



Bryan Orme

APPENDIX A: THE SBC DESIGN ALGORITHM

Step 0: Interview an initial 30 to 60 respondents using the CBC-looking questionnaire described above with a traditional (non-adaptive) random design (different version for each respondent). It is best if the design used for these initial respondents has excellent one-way and two-way frequency balance across the attributes (such as produced by Sawtooth Software's "Balanced Overlap" algorithm).

Step 1: Using all respondents' choice data collected to this point, tabulate raw frequency counts for all 2-way (between two attribute) contingency tables, representing how many times each 2-way attribute combination was chosen. After adding the constant 5 to each raw count (as a naïve prior and for robust Chi-square computation in the case of sparse data), identify the 2-way counts table that has the most significant interaction⁸ effect according to the Chi-square test (the smallest p-value). (If the raw count for the most preferred level combination within this 2-way counts table is greater than about 100, then proceed to step 3.)

Step 2: Draw (randomly) a two-attribute level combination from the most significant interaction with likelihood proportional to raw counts (let's say it's attribute 2 level 1, attribute 4 level 5). If there are more than 2 attributes in the study, use the drawn two-attribute combination as a *filter* (only count tasks where A2L1 & A4L5 are chosen) and across all respondent data collected to this point (satisfying the filter) count the frequencies for all other attributes taken independently (after adding a constant of 5 to each level count to avoid any counts of zero⁹). Make a draw for all remaining attributes with probability proportional to these filtered raw counts, thus completing the designed concept. Repeat Step 2 to draw as many concepts for this respondent as needed to design all tasks (reject any concept that is either prohibited or identical to a previously chosen concept within the same task). Field the tasks and collect data for this respondent. Go back to step 1 for the next respondent or go to step 6 if done with all data collection.

Step 3 (same approach as Step 1, but using 3-way counts to identify the most significant interaction): Using all respondents' choice data collected to this point, tabulate raw frequency counts for all 3-way (between three attribute) contingency tables, representing how many times each 3-way attribute combination was chosen. After adding the constant 5 to each count (for robust Chi-square computation), identify the 3-way counts table that has the most significant interaction effect according to the Chi-square test (the smallest p-value). If the raw count for the most preferred level combination within this 3-way counts table is greater than about 100, then proceed to step 5.

Step 4 (same approach as step 2, but leveraging 3-way counts): Draw (randomly) a three-attribute level combination from the most significant interaction with likelihood proportional to raw counts (let's say it's attribute 2 level 1, attribute 3 level 3, and attribute 4 level 5). If there are more than 3 attributes in the study, use the drawn three-attribute combination as a *filter* (only count tasks where A2L1, A3L3, & A4L5 are chosen) and across all respondent data collected to this point count the frequencies for all other attributes taken independently (after adding a constant of 5 to each level count to avoid any counts of zero). Make a draw for all remaining

⁸ To compute interaction significance via Chi-square, compute the expected frequencies based on the main effects and compare to the actual observed frequencies.

⁹ Adding a constant of 5 makes sure that every possible product concept has a non-zero chance of being selected at each stage of the adaptive design process.

attributes with probability proportional to these filtered raw counts, thus completing the designed concept. Repeat Step 4 to draw as many concepts for this respondent as needed to design all tasks (reject any concept that is either prohibited or identical to a previously chosen concept within the same task). Field the tasks and collect data for this respondent. Go back to step 3 for the next respondent or go to step 6 if done with all data collection.

Step 5: Repeat same pattern as above for the four-attribute interaction table, etc. until done with all data collection.

Notes

1. For the analysis phase, SBC could be made to handle quite extensive prohibitions between attributes taken two at a time. Two attributes involving extensive interactions could be collapsed into a single attribute (factor). Well-chosen prohibitions may help matters during the analysis since they reduce the design space and focus the respondent on relevant concepts. As for the design phase, the Chi-square interaction statistic might be computed after imputing data for missing (prohibited) cells. We could use rejection sampling to discard any prohibited combinations that we might draw for inclusion in the design.
2. SBC designs could accommodate *a priori* defined segments. For example, if you believed that males and females had strong differences in preference and wanted to design optimal products for males vs. females within the same questionnaire, you would simply include a filter on the counts in each of the design steps based on male/female. Because the data would be now split in two, the initial sample size prior to graduating to the adaptive design would need to be doubled. Total sample size would also need to be twice as large (if requiring equal precision for segment-based product optimization across two evenly sized segments).
3. Counting analysis and drawing new products proportional to prior choice frequencies is so rapid that each respondent should have very little wait time for each task to be generated. The initial identification of the most significant interaction effect should be done just once per respondent and retained for generating all subsequent tasks for that respondent.
4. Although SBC designs deviate significantly from level balance and orthogonality, the design supports standard logit, HB, or latent class estimation. To make this possible, when designing each choice task, a few (such as 20%) of the product concepts should be drawn purely randomly (not according to prior preferences). This provides adequate design efficiency for running MNL (aggregate or disaggregate) on the data. Drawing a few purely random concepts per task is also a way to reduce the likelihood that biased early respondents could hinder the ability of the solution to converge upon globally optimal product concepts.

OPTIMAL PRODUCT DESIGN BY SEQUENTIAL EXPERIMENTS

MINGYU JOO

UC RIVERSIDE

MICHAEL L. THOMPSON

PROCTER & GAMBLE

GREG M. ALLENBY

OHIO STATE UNIVERSITY

ABSTRACT

Optimal product design is challenged by the presence of attributes with many levels that are thought to interact with each other. Product and package colors, tag lines, styles and visuals are examples of attributes with a “flat” space of attribute-levels that is difficult to parameterize. Compounding this problem is the interest to identify interactive effects among attribute-levels, such as certain color combinations and messaging strategies that are thought to increase sales. This paper introduces a general framework for identifying these high-dimensional interactions in the context of a sequential experiment.

INTRODUCTION

Optimal product and package design relies on identifying interactions among attributes and their levels that drive sales. At the heart of any effort to uncover interactions are three things: 1) a sequential framework for learning about interactive effects; 2) a model of the data that is sufficiently informative about the presence of interactive effects so they are not overlooked, i.e., that reduce errors of omission; and 3) a flexible method of shutting-down, or eliminating effects that are small or insignificant so that errors of commission are also not present.

In this paper we propose use of a sequential experiment that facilitates learning about interactive effects and using this information to select the next-best set of product profiles to test in an analysis. This allows us to learn about important attributes and potential interactive effects from scratch, where the data drives the results and not prior knowledge of the researcher. The use of sequential experiments to cover high-dimensional space of possible effects is used in fields ranging from medical testing (e.g., Bartroff, Lai, Shih 2013) to engineering optimization (e.g., Jones, Schonlau, Welch 1998).

An informative model is the second necessary ingredient to obtaining a viable procedure for product design. Over the course of the last 25 years, researchers in conjoint analysis and demand modeling have been incorporating heterogeneous effects with the use of hierarchical Bayes (Rossi, Allenby, McCullough 2005) methods. These models, however, are often estimated with sparse data and do not include interactive effects. Heterogeneous effects, therefore, work against the discovery of interactions because these models are heavily parameterized. In this paper we rely on an aggregate model of demand in which the effects of heterogeneous coefficients are minimized.

Third, we avoid the identification of false-positive interactive effects using Bayesian variable selection (George and McCullough 1993) methods. In our analysis we find that Bayesian variable selection out-performs other methods and minimizes the detection of false-positives.

We develop our model within the context of a package design problem faced by a leading consumer-goods manufacturer. Through simulation, we find that five rounds of a sequential experiment are expected to be sufficient for interaction detection. We then design a study and apply our model to data in which respondents identify the best package design from a set of alternatives. The best package design from our sequential experiment is then compared to alternative designs that were deemed best from alternative methodologies in a second study. The results from this best-of-class comparison favors our proposed method in comparison to other methods used by the manufacturer.

The next section introduces our proposed design criterion that favors design points, or concepts for inclusion in the study having the greatest likelihood of improving expected sales beyond the current best design. This criterion is expressed as the upper tail of the expected predictive distribution, given information about what is currently known about the model parameters, or part-worths. As the experiment proceeds, new data are collected and used to form revised estimates of the model parameters, which are then used to form revised estimates of the predictive distribution of alternative design points. The sequential nature of the experiment effectively searches over potential product designs that have the greatest chance of improving sales and avoids testing combinations of features expected to have low sales. Thus, our design criterion is different from traditional design criteria that seeks to optimally estimate all model parameters.

MODEL DEVELOPMENT

We begin the discussion of our methodology by first describing the model used to analyze the data. We show that the aggregate model can be easily linearized and the effects of heterogeneous effects across respondents are ignorable for evaluating which design consumers prefer most. We then describe our design criterion. Details of our method are described in Joo et al. (2018).

The goal of our analysis is to find the product design that maximizes sales to the firm. We do this using a model of aggregate shares from an experiment where respondents are asked to indicate the best design from a series of alternatives. We employ an aggregate share model for analysis:

$$S_{jq} = \frac{\exp(u(\mathbf{X}_j; \boldsymbol{\beta}))}{1 + \sum_{j' \in J_q} \exp(u(\mathbf{X}_{j'}; \boldsymbol{\beta}))} \text{ and } S_{0q} = \frac{1}{1 + \sum_{j' \in J_q} \exp(u(\mathbf{X}_{j'}; \boldsymbol{\beta}))},$$

where S_{jq} is the share of product profile j in question q , and S_{0q} is the share of a common outside option in question q , which can be the current design on the market, a vanilla design profile, or no choice. J_q includes all product profiles j' tested in question q . The utility of product profile j in question q , $u(\mathbf{X}_j; \boldsymbol{\beta})$, is determined by a linear combination of the design attributes and the vector of part-worths. Taking the log-odds of S_{jq} and S_{0q} results in a linearized version of the logit model

$$\ln S_{jq} - \ln S_{0q} = u(\mathbf{X}_j; \boldsymbol{\beta}) = \mathbf{X}_j \boldsymbol{\beta} + \xi_{jq},$$

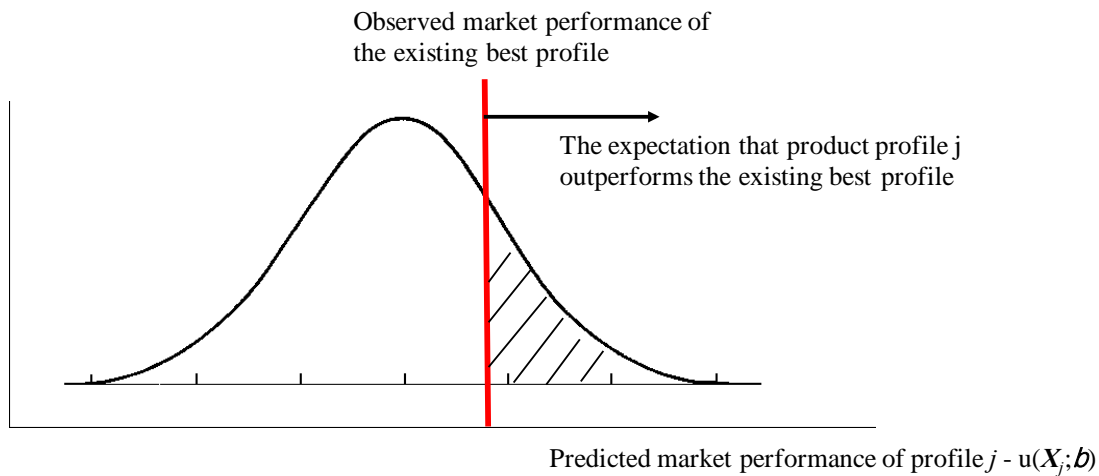
where ξ_{jq} is a mean-zero, normal error term specific to product profile j in question q . We note that \mathbf{X}_j includes main-effect and interaction terms that can be very high dimensional. Consider, for example, the dimensionality of the design space for the color of a bottle cap and the color of lettering on the bottle. The dimensionality of the design space for just 10 colors for each is at least 300 when considering the two-way interactions alone. This illustrates the need to work with an aggregate model of demand, and the difficulty in exploring interactive effects with individual-

level demand models where the number of data points is typically constrained to be less than 20 observations.

An advantage of working with an aggregate-level linear model is that the effects of heterogeneous coefficients can be ignored when the goal is to estimate the average utility of the product profiles. The average utility is appropriate in product and package design contexts when respondent-level customization of the product is not being considered (Theil 1965). Thus, the average effect of the attribute-levels is directly estimable from the linearized model, and can be used to predict the aggregate market-level response to the product or package design.

The second element of our method is the design criterion that iterates towards the best design points, or product configurations, to include in the experiment. We assume that, at any point in time in the sequential experiment, there exists a product configuration thought to be the best given the information currently available from the data. Given this prior information about the part-worths, it is possible to obtain the predictive distribution for any new product configuration. Our proposed design criterion is the probability that the new configuration is better than the existing best, which is represented as the area to the right of the vertical bar in Figure 1. The designs that are most promising, measured as having the greatest probability mass to the right of the best-performing design observed in the current round, are the designs to include in the next round of the experiment.

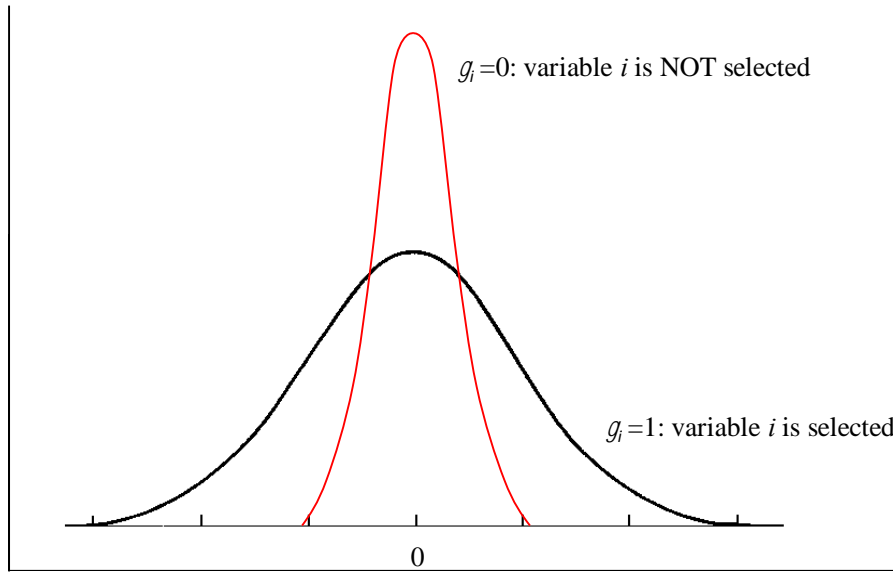
Figure 1. Proposed Criteria



Our proposed criterion differs from traditional experiment design criteria, such as D-optimality (Box, Hunter, Hunter 1978) which seeks to minimize the variance of all model parameters. Our design criterion favors product configurations with a high likelihood of improving upon the best configuration already tested. Thus, it identifies design points that simultaneously increases “learning” and “earning” (Schonlau et al. 1997, 1998).

The third element of our proposed method is the use of Bayesian variable selection (George and McCulloch 1993) to identify non-zero effects in the model. We find that Bayesian methods are good at minimizing errors of commission (Type II errors). Bayesian variable selection works by specifying a “mixture” prior distribution on the model parameters that give appreciable probability mass to the regression coefficient (β_i) equal to zero or having no effect in the model.

Figure 2. Mixture Prior for Variable Selection



Thus, there is some probability $P(\gamma_i=0)$ that the β_i is very close to zero, and probability $(P(\gamma_i=1)=1-P(\gamma_i=0))$ that β_i is described by a prior distribution with much larger variance. Bayesian variable selection produces an estimate of $P(\gamma_i=0)$ for each regression coefficient, allowing for a sharper assessment of which regression coefficients are significantly different from zero and which are not.

SIMULATION STUDY

We investigate the performance of our proposed method encompassing 1) the aggregate share model, 2) the design criterion and 3) Bayesian variable selection in a sequential experiment setting where one of four product concepts is chosen in twenty-five choice tasks for each wave of the experiment. The choice tasks also have a common product configuration, which we refer to as the “outside option” so that the aggregate share model is linearized consistently across the choice tasks. A researcher can use either a “vanilla” design profile or the current product design on the market as a common outside option. One hundred simulated respondents indicate their most preferred configuration, and these choices are aggregated to form the shares used in estimation. Thus, each round of the experiment is capable of evaluating $25 \times 4 = 100$ product configurations, and we examined the performance of the method over five waves of the sequential experiment.

Each product profile is a combination of four attributes, and these attributes have design candidates of 5, 8, 11, and 12 levels each. The number of attributes is set to be similar to the collaborating firm’s R&D project. The true part-worth preference parameters are designed in a way that the true best design concept involves two-way interactions. The dummy coding of all design attributes leads to 32 main effects and 369 two-way interaction effects. The product design problem includes 5,279 candidate product profiles. Therefore, the simulation tests the performance of the proposed framework where the number of product profiles tests (500) is similar to the number of parameters (401).

We compare our method to the polyhedral method for adaptive choice-based conjoint analysis proposed by Toubia et al. (2004). The polyhedral method is one of the earliest developments in machine-learning based question-selection methods in conjoint analysis, and is shown to be very

efficient in reducing uncertainty in part-worth parameter estimation relative to traditional conjoint analysis methods. The polyhedral method iteratively selects a respondent-level choice set after a respondent completes each choice task. It aims to minimize errors around all parameters, not selectively around more positive ones.

Table 1 compares the performance of the proposed framework to the polyhedral method for highly preferred design profiles. Reported is the correlation between the true and predicted utilities for the top 1% of the profiles. The results of the simulation study indicate that the proposed design criterion is effective at identifying the most preferred product concepts.

Table 1. Correlations Between True and Predicted Utilities Out of Top 1% of Profiles

Profiles	Proposed criteria	Polyhedral method
Top 100 (0.2%)	0.638	0.263
Top 200 (0.4%)	0.635	0.460
Top 300 (0.6%)	0.688	0.575
Top 400 (0.8%)	0.697	0.650
Top 500 (1.0%)	0.668	0.702

EMPIRICAL APPLICATION

The proposed framework is applied to a package design project for a consumer packaged good of a leading consumer products manufacturer. The manufacturer's goal is to develop the optimal design in the presence of high-dimensional and sparse parameter space, avoiding a poor combination of the best levels of main effects. The expected improvement criterion is implemented to search for the most promising design concepts with highest potential, using the parameter estimates by the Bayesian variable selection method. The study is conducted online with high-resolution images of hypothetical product packages.

The product package consists of four design attributes including visual image of the product, claim statement of key features, name of materials used in the product, and sub-brand name, as described in Table 2. The design element of the main brand name is fixed, so that is not included in the design experiment. The manufacturer selects the candidate attribute levels using domain knowledge including 12 product images, 11 claim statements, 6 materials, and 12 sub-brand names. One level from each attribute is considered to be as baseline with part-worth preferences of zero for identification purpose. The baseline attributes construct the common outside option in the experiment.

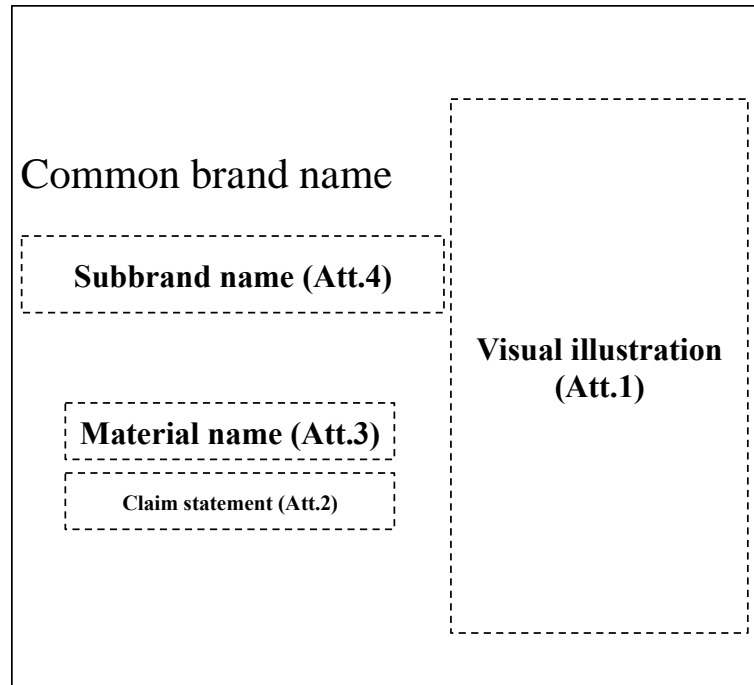
Table 2. Description of Attributes

Attribute codes	Numbers of levels	Description	Space allotted
Att.1	12	Visual illustration of the product	High
Att.2	11	Claim statement of the key strength	Low
Att.3	6	Name of material	Medium
Att.4	12	Sub-brand name	Medium

Figure 3 illustrates the design elements of the product package. Attribute 1 (visual illustration) is the largest design element of the package in addition to the common main brand, so it is highly visible to respondents. Attribute 2 (claim statement) is at the bottom of the package with a small font, but includes information important to respondents. Attribute 3 (material) is placed right above the claim statement with a larger font. Attribute 4 (sub-brand name) is placed right below the common brand name with a similar sized font as Attribute 3. The design

attributes' visibility may correlate with the size, but it does not necessarily reflect the importance of information.

Figure 3. Location of Design Attributes in the Package



The number of rounds in the sequential experiment is predetermined as five, considering the size of data required for accurate parameter estimation and the manufacturer's typical budget limitation for R&D projects. About 450 respondents per round participated from the U.S. and the U.K. as in Table 4. All participants confirmed that they are active users of the focal product category through screening questions.

Table 4. Summary of Sample Sizes in Each Round of Experiment

	Round 1	Round 2	Round 3	Round 4	Round 5
U.S.	228	223	237	217	233
U.K.	224	227	214	233	218
Total	452	450	451	450	451

Respondents receive three hypothetical design concepts and one common outside option in each question. They are requested to select their favorite design concepts out of four alternatives, as described in Figure 4. The displaying order of the four design concepts is randomized for each respondent to avoid any location effect. Respondents can enlarge the pictures of each of the given package design concepts to the full screen mode for evaluation.

Figure 4. Screen Layout for the Conjoint Experiment

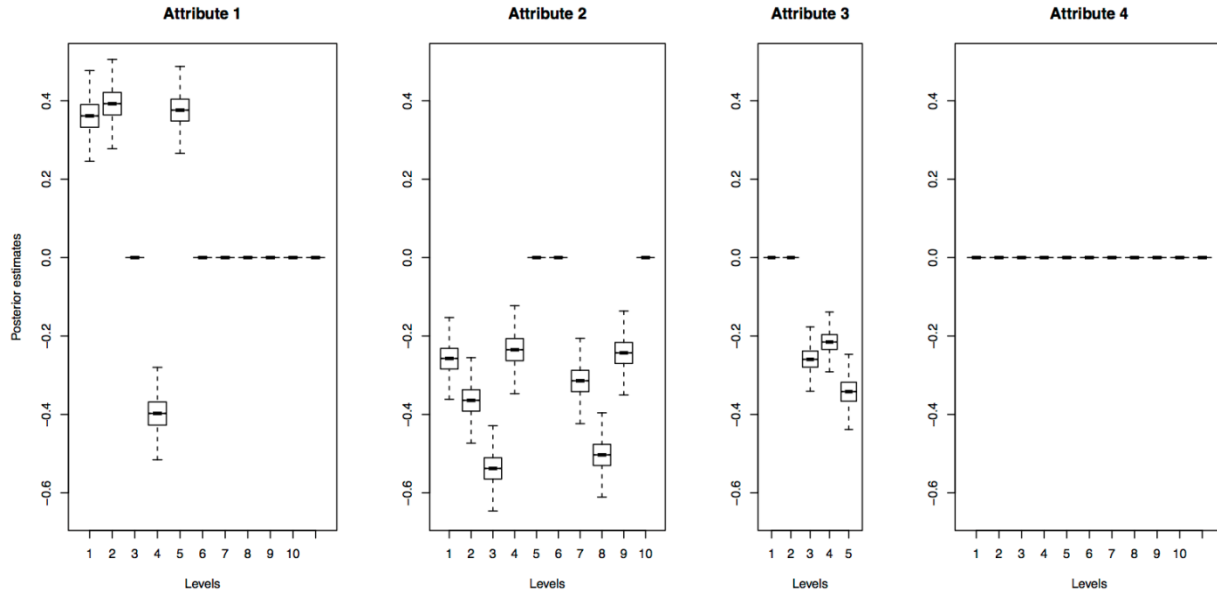
Please select the [*brand name*] package below that you would be **Most Likely** to purchase.

Package design alternative 1	Package design alternative 2
<input checked="" type="radio"/>	<input type="radio"/>
Package design alternative 3	Package design common outside option
<input type="radio"/>	<input type="radio"/>

Each respondent provided answers to 23 choice tasks in each round. They are randomly split into five groups receiving the different product profiles for evaluation. The number of alternatives tested in one round is 345 ($=23 \times 3 \times 5$). The product profiles to be tested in the first round are selected by classical criteria. The part-worth preference parameters are estimated at the end of data collection in each round (t) using the Bayesian variable selection method, and the design profiles to be tested in the next round ($t+1$) are determined by the proposed expected improvement criterion conditional on the parameter estimates. The same procedure is iterated until the end of the fifth round.

Figure 5 presents the summary of posterior estimates for main effect parameters using the data from all five rounds of experiment. Attribute 1 (visual illustration) is the most important design attribute in terms of the variation across levels, while attribute 4 (sub-brand name) does not affect preferences. The visual element is the largest part in the package design, so it may attract the highest level of attention from respondents.

Figure 5. Posterior Estimates for Main-Effect Part-Worth Parameters



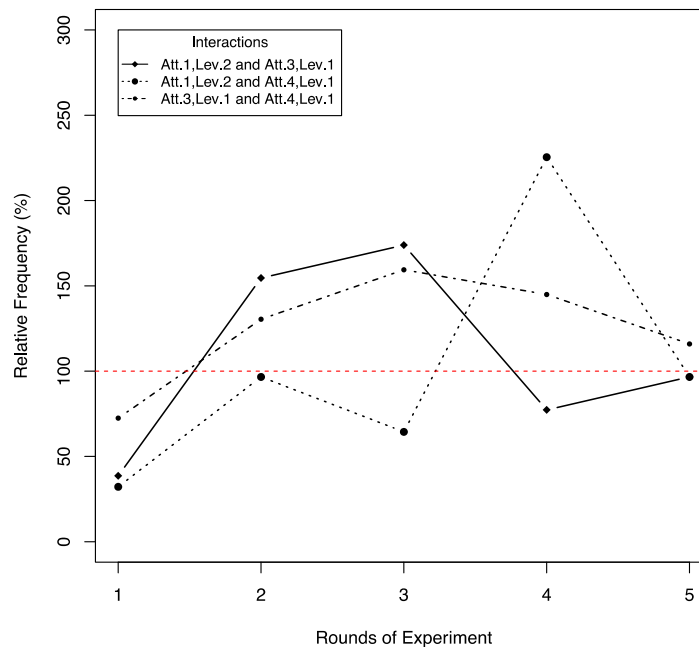
DISCUSSION

The empirical results indicate that the optimal design is affected by the presence of interactive effects among the attributes. Two natural questions that arise are 1) whether the proposed criterion leads to the evaluation of design concepts with more appropriate interactions, and 2) whether the optimal design by the proposed framework is actually preferred to designs suggested by other methods.

The following discussion first examines the effects of selection criteria on the frequencies of appropriate interactions evaluated in the experiment. Then, it presents a validation task to evaluate the performance of the proposed framework relative to other popular methods in practice. For validation, the collaborating firm implemented three more R&D experiments using competing methods, and we conducted another separate survey to evaluate three best profiles suggested by the competing methods against the one by our proposed framework.

Figure 6 presents the observed relative frequencies of key interaction effects appearing in the product profiles that are evaluated in the five rounds of experiment. The interaction effects listed are those that appeared in the most preferred design concept. The dotted line indicates the expected frequency of evaluation, if candidates are randomly selected without applying proposed adaptive selection criteria. If the key interaction effects are more frequently selected to evaluate in the earlier rounds, it is highly likely to find the best design concept without too many rounds of iterations.

Figure 6. Relative Observed Frequency of Important Interaction Effects



The relative frequency plot illustrates that the proposed criterion allocates the limited number of questions in a more efficient way to search for the potentially best combinations of design concepts. The respondents are directly exposed to the combinations with appropriate interaction effects. This confirms that the important interactive effects between attributes are less likely to be omitted in the prediction of optimal design, as respondents make head-to-head comparisons among highly preferred combinations.

The empirical application reported above was conducted as part of a large-scale R&D project by the manufacturer, which consists of four separate product design experiments including our proposed framework. The manufacturer has relied for a long time on methodologies offered by commercial vendors, Nielson's optimizer with evolutionary genetic algorithm and Sawtooth Software's choice based conjoint (CBC) experiment. They also adopted and developed a machine-learning based query-selection method, called optimal Bayesian recommendation set (Viappiani and Boutilier 2010). Therefore, the R&D project produced four different product profiles recommended by each method—the proposed one, genetic algorithm, standard CBC, and Bayesian recommendation set.

An additional validation survey was conducted to directly compare the proposed optimal profile with the other three profiles created by the competing methods. A new set of individuals was selected and responded to one choice task of their favorite design among the four profiles and a no-choice option.

The three other design experiments were conducted for exactly the same product package described above. They were implemented under supervision of the collaborating firm with software providers, and we have limited information on details of implementation except for the final outcome. Therefore, we briefly describe the three benchmark methods at a conceptual level.

Nielson's genetic algorithm. Nielson's optimizer adaptively searches for the best product designs at the individual level using interactive genetic algorithm based on Malek (2001). The genetic algorithm is a heuristic approach to mimic nature's evolutionary process, where superior

ones eventually survive (Balakrishnan and Jacob 1996). The questionnaire starts with a random initial set of product profiles. Subsequent sets of questions present superior offspring of product profiles in the previous round, i.e., combination of preferred attribute levels. The algorithm allows mutation in general for exploration purposes. Empirical studies have shown that the outcome of genetic algorithm is often close to optimal and outperformed other existing heuristic methods (Balakrishnan and Jacob 1996), as multiple iterations of evolution improve the fitness of the outcome.

Sawtooth Software CBC. The choice-based conjoint (CBC) method uses a standard hierarchical Bayes conjoint model provided by Sawtooth Software. They offer 30 different blocks of 20 predetermined choice-tasks each, using a “Balanced Overlap” experimental design. Respondents first build their own designs using a graphical configurator, then are randomly assigned to one questionnaire block out of 30 sets. The responses out of 20 questions per individual are analyzed by hierarchical Bayesian method (main effects only, without using the configurator information), so the part-worth preference parameters are estimated at the individual level accounting for heterogeneity.

Bayesian recommendation set. A machine-learning algorithm searching for optimal recommendation sets (Viappiani and Boutilier 2010) is adopted by the researchers of the manufacturer. Viappiani and Boutilier (2010) show that the myopically optimal choice set in an adaptive experiment is equivalent to the optimal recommendation set of the same size, i.e., a set of product profiles that maximizes the respondent’s expected utility. In the sequential process, it presents a set of product profiles to test in the next round that maximizes their sum of expected utilities using part-worth preference parameter estimates in the previous round. The part-worth parameters are estimated by hierarchical Bayesian method accounting for heterogeneity of individual respondents.

The four separate experiments including the proposed method and three benchmark methods described above result in four different optimal design profiles for the identical product package. All four methods predicted four different best profiles with internal validity according to preference estimates from each model, but separate results are not able to present external validity. Therefore, we conduct a separate validation survey to compare four different design profiles out of different experimental methods.

Participants include 523 individuals from the U.S. ($n=266$) and the U.K. ($n=257$) and are active users of the focal product category. All respondents receive one question of choosing their favorite design concept out of four product profiles including the proposed one and no choice option. The orders of four design concepts are randomized to avoid location effects.

Table 5 presents the shares of the four design concepts generated by different methods in the manufacturer’s R&D project. The optimal design created by the proposed framework is the most preferred design concept out of the four product profiles, each of which is predicted as the best design by different methods. The proposed design lifts the observed share by 8% relative to the design by genetic algorithm, and by 14% and 52% relative to the standard choice based conjoint and the machine learning method, respectively. We note that all three benchmark methods fully controlled respondents’ heterogeneity, and especially genetic algorithm is provided at a very high cost to the manufacturer. Though we are not able to offer the market share prediction based on this result, the observed improvement is potentially significant considering that the manufacturer’s revenue per brand is over \$1 billion on average.

Table 5. Validation Experiment

	Choice frequency	Proportion	Relative share lift by the proposed design
Proposed method	140	26.8%	8% 14% 52%
Genetic algorithm	130	24.9%	
Standard CBC	123	23.5%	
Optimal recommendation set	92	17.6%	
No-choice	38	7.3%	
Total	523	100.0%	

The validation results show that the proposed framework identifies the optimal product profile by prioritizing appropriate combinations of attributes in the sequential test. The standard choice-based conjoint analysis relies on a classical experimental design, which frequently produces a main-effect design without interactions. Genetic algorithm (Malek 2001) and optimal Bayesian recommendation set (Viappiani and Boutilier 2010) are designed to overcome such problems, but they are sensitive to the initial seed with limited exploration and rely on heuristic comparison in a subset of product profiles. The results confirm that the share model used in our proposed framework is suitable for identifying market-share maximizing design concepts.

CONCLUSION

This paper proposes a new approach to optimal product design in high dimensions using sequential experiments. Product profiles are prioritized for inclusion if they can improve on the outcome of the current best design. The expected improvement criterion is operationalized by an integration of upper tail in the posterior distribution of aggregate market share. A stochastic search variable selection method reduces the dimensionality of the model by selecting relevant variables. We demonstrate that the proposed framework identifies the best design in a large-scale R&D project conducted by a major packaged goods company.

Our modeling framework can be applied to many high-dimensional design settings, such as identifying brand logos, optimal advertising campaigns, etc. It can also be applied to other R&D projects with horizontal variation in the attribute levels. The proposed framework is especially effective when the design attributes contain a large number of levels, and the evaluation of all potential candidates is infeasible.



Mingyu Joo



Michael L. Thompson



Greg M. Allenby

REFERENCES

- Balakrishnan, P.V. (Sundar), Varghese S. Jacob. 1996. Genetic algorithms for product design. *Management Science*. 42(8), 1105–1117.
- Bartroff, Jay, Tze Leung Lai, Mei-Chiung Shih. 2013. Sequential Experimentation in Clinical Trials: Design and Analysis. Springer.
- Box, George E.P., William G. Hunter, J. Stuart Hunter. 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley and Sons, Inc.
- George, Edward I., Robert E. McCulloch. 1993. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*. 88(423), 881–889.
- Jones, Donald R., Matthias Schonlau, William J. Welch. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*. 13, 455–492.
- Joo, Mingyu, Michael L. Thompson, Greg M. Allenby. 2018. Optimal Product Design by Sequential Experiments in High Dimensions. *Management Science*. Forthcoming.
- Malek, Kamal M. 2001. Analytical and Interpretive Practices in Design and New Product Development: Evidence from the Automobile Industry. Ph.D. thesis, Massachusetts Institute of Technology.
- Rossi, Peter E., Greg M. Allenby, Robert E. McCulloch. 2005. *Bayesian Statistics and Marketing*. John Wiley and Sons Ltd.
- Schonlau, Matthias, William J. Welch, Donald R. Jones. 1997. A data-analytic approach to Bayesian global optimization. *American Statistical Association Proceedings, Section of Physical Engineering Sciences*. 186–191.
- Schonlau, Matthias, William J. Welch, Donald R. Jones. 1998. Global versus local search in constrained optimization of computer models. *New Developments and Applications in Experimental Design*. 34, 11–25.
- Theil, Henri. 1965. *Linear Aggregation of Economic Relations*. North Holland Publishing Company.
- Toubia, Olivier, John R. Houser, Duncan I. Simester. 2004. Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis. *Journal of Marketing Research*. 41(1), 116–131.
- Viappiani, Paolo, Craig Boutilier. 2010. Optimal Bayesian Recommendation Sets and Myopically Optimal Choice Query Sets. *Advances in Neural Information Processing Systems*. 23, 2352–2360.

SEGMENTATION ANALYSIS VIA NON-NEGATIVE MATRIX FACTORIZATION

MICHAEL PATTERSON

JACKIE GUTHART

CURTIS FRAZIER

RADIUS GLOBAL MARKET RESEARCH

ABSTRACT

Non-Negative Matrix Factorization (NMF) is a relatively new technique that allows for the simultaneous segmentation of individuals and “factoring” of variables. This paper will introduce NMF and compare its performance relative to standard segmentation approaches (K-means Clustering, Latent Class Analysis, Hierarchical Clustering) using both simulated data and data from an actual study. The results from our analysis show that NMF performs very well, especially in the case of highly correlated datasets.

INTRODUCTION

When conducting a segmentation analysis utilizing survey data (e.g., via needs or attitudes) researchers often find that many of the basis variables are correlated to varying degrees. In cases of high to very high correlations this can produce a potentially biased solution since the correlated variables may unduly influence the solution. This is true for traditional techniques such as K-means Clustering as well as more advanced approaches such as Latent Class Analysis (LCA).

There are several different approaches for dealing with datasets containing multicollinearity such as grouping together items, either as factor scores via Principal Components Analysis or as composite variables (either simple averages or weighted averages), or by removing somewhat redundant items altogether, retaining only the most representative items. However, these are not necessarily statistically optimal solutions and in some cases can be somewhat arbitrary approaches.

Another fairly frequently encountered issue in segmentation is the analysis of sparse data as well as binary data. In the case of sparse data, when individuals are asked to indicate which activities they have engaged in, products they have purchased, sources consulted, etc. we may have only a relatively few number of selections per item, particularly when there are a fairly large number of items. In our experience we have found that K-means Clustering with Euclidean distances in particular is not well suited to the analysis of data that are sparse and/or binary.

In this paper, we will introduce a relatively new analytic approach, Non-Negative Matrix Factorization (NMF) that addresses these common issues and provides much greater insights into the data.

An NMF analysis simultaneously takes into account the relationship between the segmentation basis variables while also forming the segments. That is, items are grouped together in “factors” or latent variables, at the same time that individual respondents are grouped together in segments.

Below, we explore how NMF compares to traditional segmentation approaches such as K-means Clustering, LCA, and Hierarchical Clustering, as well as ensembles, based on simulated data with known properties (e.g., little/moderate/high multicollinearity, little/moderate/high sparseness). Hit or concordance rates (actual segment classification vs. recovered) will be compared across the techniques.

In addition, we demonstrate the use of NMF vs. other approaches utilizing data from an actual client study to show the enhanced interpretability of NMF. The R code that we utilized is available, allowing others to replicate the simulated results, at http://www.sawtoothsoftware.com/download/Patterson_Guthart_Frazier_2018.zip.

POTENTIAL SEGMENTATION ISSUES

A typical segmentation analysis uses various inputs such as scaled attitudes, scaled needs, behaviors, or utilities derived from MaxDiff.

Often we find it can be helpful to process the data to introduce greater discrimination. For example, we might recode items in the top 10 as binary (“1” vs. “0”) or code Top-2-box scores as binary. However, when including behavioral data (e.g., activities conducted, products purchased, etc.) this can lead to sparse data that contains many “0”s and few “1”s (where the presence of a behavior is coded as a “1”). This also holds true when transforming data into binary scores, although the analyst has much more control in that case (e.g., rather than top 10, the analyst might code top 20 in order to increase the frequencies of “1”s and thereby make the data less sparse). In any case, running a segmentation analysis of binary data can be problematic, particularly with techniques such as K-means Clustering, which works best with scaled data.

Another issue frequently encountered when using scaled values is correlated items (particularly when there is only one set of highly correlated items). When items are moderately/highly correlated, those variables may exert more influence on the solution.

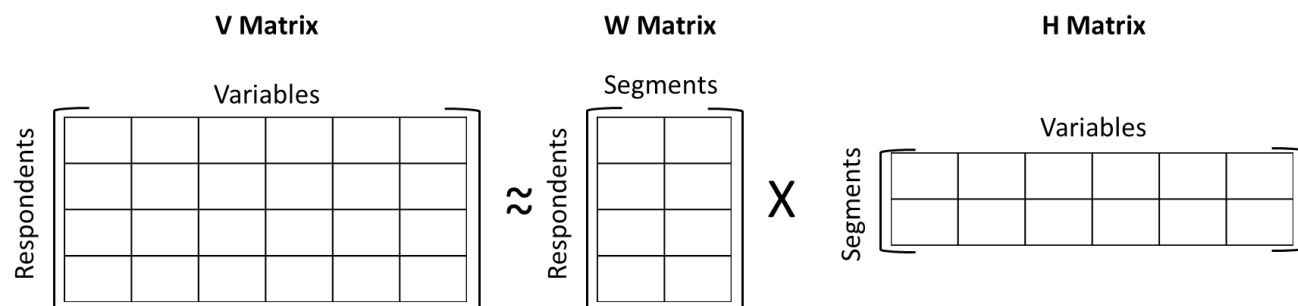
*“High correlation among clustering variables can be problematic because it may overweight one or more underlying constructs”—
Ketchen and Shook (1996)*

POSSIBLE SOLUTIONS

Some commonly used solutions to the aforementioned potential segmentation issues include: using factor scores, using composite variables (that combine variables), or using only one variable from each correlated set. Another potential solution for dealing with both sparse data and correlated attributes is Non-Negative Matrix Factorization—also referred to as NMF (Lee and Seung, 1999).

NMF analysis simultaneously takes into account the relationship between the segmentation basis variables while also forming the segments. That is, items are grouped together in “factors” or latent variables, while respondents are simultaneously grouped together in segments. NMF can use any non-negative data type, including binary data such as associations, dichotomized ratings, activities, etc.

The principle behind the approach is to factor a given data matrix (traditionally called “V”) into two lower dimensional matrices (W and H) such that none of the three matrices contains any negative items (hence the term non-negative):



The resulting matrices contain the “loadings” of Respondents on Segments (Matrix **W**) along with the loadings of Variables on Segments (Matrix **H**). In this way, we are able to simultaneously group individuals into segments (based on the highest loading) along with variables into “factors” providing much greater insights into the structure of the data. Various algorithms (often Gradient Descent) exist to identify the Matrices **W** and **H** that minimize $|V - WH|^2$ (i.e., the factoring is not exact).

Non-Negative Matrix Factorization is one approach to Matrix Decomposition (others include Singular Value Decomposition and Principal Components Analysis) and has been used to study such areas as: facial and image recognition, recommender systems, text mining, acoustic signal processing, financial and stock trading data, and retweeting behavior. It is relatively new in application to market research segmentation.

SIMULATED DATA

In our simulation research we compare NMF with various segmentation algorithms. To do this, we looked at six relatively common types of datasets, each with known properties and known segments. We generated 100 random versions of each type of dataset, with 1,000 respondents in each. We analyze the datasets via four algorithms: Non-Negative Matrix Factorization, Latent Class Mixture Model, K-means Clustering, and Hierarchical Clustering. Then we compare the recovered segments to the known segments—the concordance or “hit” rate.

The six types of datasets differ in terms of their level of sparseness and in the correlations among related items. Sparseness is defined as the probability that a variable is coded as “1” vs. “0.” For the correlations, groups of items were designated as belonging together (similar to a factor) and then correlations among these common items were established. The correlations below show the average correlation among these “related” items.

Dataset Type	Description	Level of Sparseness	Correlations Among Similar Items
1	Very sparse, no corrs	5%	~0.00
2	Moderate sparse, low corrs	33%	0.28
3	Moderate sparse, moderate corrs	33%	0.51
4	Moderate sparse, high corrs	33%	0.80
5	Low sparse, moderate corrs	50%	0.53
6	Low sparse, high corrs	50%	0.80

A small number of records from two example datasets are shown below:

Dataset 1—Very Sparse (~5% of Values Are “1”), Virtually No Correlations Among Items

Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18	Item19	Item20	Actual Segment
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	4
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4

Dataset 6—Low Sparse (50% of Values Are “1”), High Correlations (~0.8 Among Related Items)

Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18	Item19	Item20	Actual Segment
1	1	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	1	0	1
1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	0	0	1
0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2
0	0	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	3
0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	3
0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0	4
0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	4

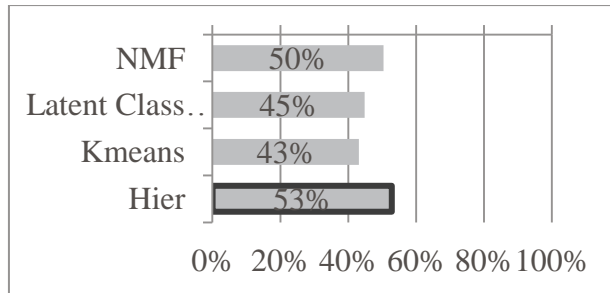
SIMULATED DATA RESULTS

The charts below show how well each method recovered the underlying true segments in each type of dataset. The percentages are the percent of respondents correctly placed.

Low Correlation

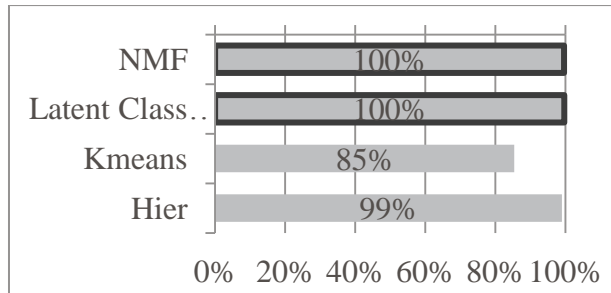
Dataset 1

Very Sparse, No Correlation



Dataset 2

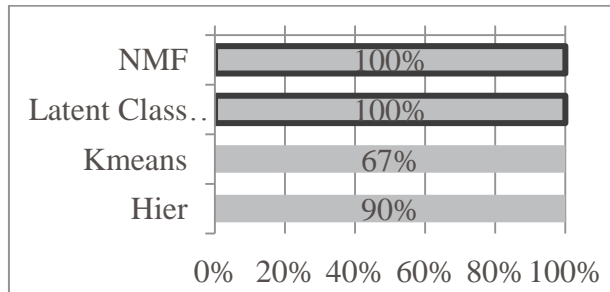
Moderate Sparseness, Low Correlation



Moderate Correlation

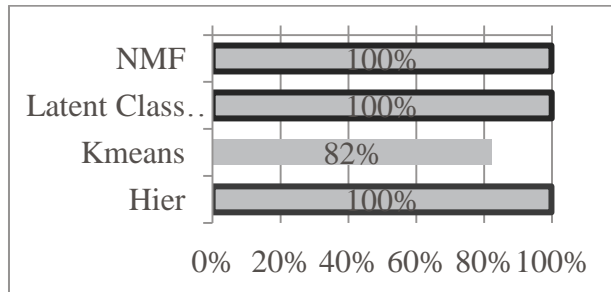
Dataset 3

Moderate Sparseness, Moderate Correlation



Dataset 5

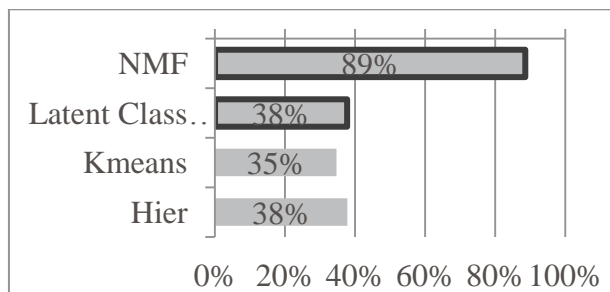
Low Sparse, Moderate Correlation



High Correlation

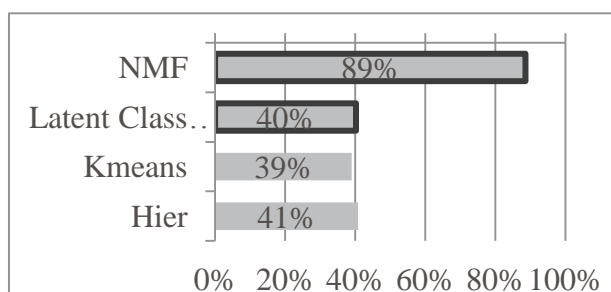
Dataset 4

Moderate Sparseness, High Correlation



Dataset 6

Low Sparseness, High Correlation



Among the four approaches that were examined, NMF is the overall top-performing method we tested. The only time it was not beat is in the case with very sparse data, where hierarchical clustering beat it by 3 percentage points (NMF: 50% vs. Hierarchical: 53%). NMF significantly outperforms the other methods in the highly correlated datasets—having over twice as good a hit rate compared to all other tested methods. As suspected, K-means Clustering performs the worst across all simulated datasets.

All results were analyzed via MANOVAs and ANOVAs and all were found to be highly statistically significant.

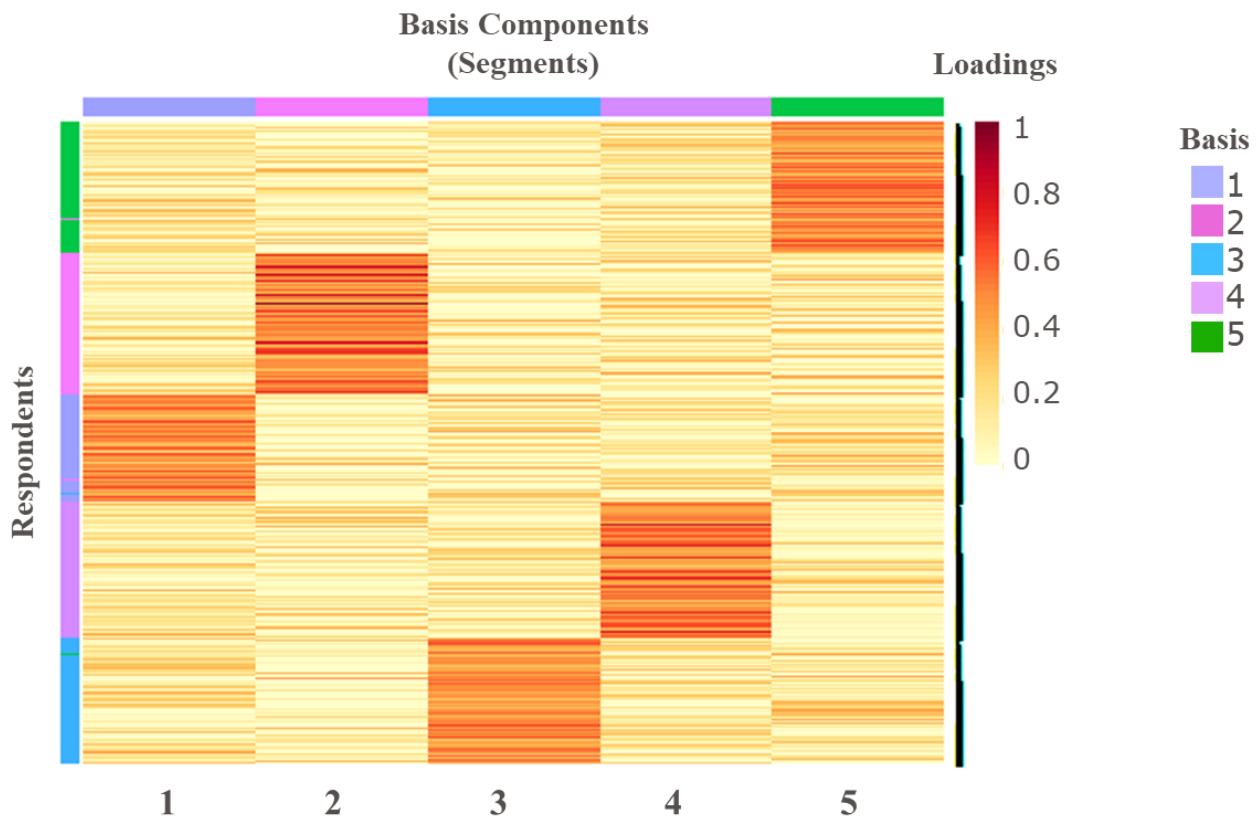
CASE STUDY

In addition to looking at simulated data, we also looked at an actual case study from the water filtration industry. Here the objective was to unbundle the market and understand the different consumer segments. A total of about 2,800 general-population consumers were surveyed online. In the survey, respondents completed two MaxDiff exercises. The first had 28 items about the category (e.g., makes my water safe, improves taste) and the second had 32 items about the product (e.g., it's effective, it's affordable). These two MaxDiff exercises were used as the basis variables in the segmentation.

For the analysis, we examined both the raw MaxDiff utilities and recoded, binary MaxDiff data (since as mentioned before we often find that this works well). For the recoding, we took the top 10 items from each exercise and coded those as a 1, and the rest as a 0. In this example the data could be classified as moderately sparse since we had 20 items out of 60 coded as 1 which is equivalent to 33% sparseness.

Example Output

The heat map below shows the Matrix W, which represents respondent loadings on segments. A respondent is assigned to a segment based on their highest loading. The Basis components at the top represent the segments, and the respondents are represented in the rows. The darker the color, the higher the respondent loads on that segment. In our experience running NMF, these clear groupings are typical.



A nice advantage of NMF is that the standard output gives additional insight about how the attributes group together. Below is Matrix H, which shows the loadings of variables on the segments.

The interpretation is that the associated items group together, similar to factor analysis. Large loadings of an item on a segment indicate that the segment is strongly defined by that attribute.

	Basic Value Shopper	Turnkey Investor	Family Guardian	Informed Performance Seeker	Brand Validator
Is affordable	1161	138	267	16	0
Does not require frequent maintenance	1040	99	205	0	0
Offers a warranty or guarantee	950	187	167	134	0
Removes contaminants	943	122	129	615	0
Maintenance is not too expensive	902	156	150	0	13
Is the most effective solution	880	89	78	628	0
Makes my water taste great	840	93	0	506	761
Makes my water safe to drink	769	58	374	518	743
Removes odor/smell from my water	726	44	0	258	699
Helps me protect myself/family	682	0	515	509	672
Gives me peace of mind and confidence	637	0	327	536	605
Reassures me that I've done all I can	601	0	359	534	563
Good flow rate of water coming out	561	135	173	0	126
Encourages me to drink more water	464	0	0	296	592
Extends the life of my appliances	168	550	0	0	12
Saves on energy	66	525	0	0	0
Saves me money	296	504	0	0	126
Is better for my skin and hair	289	414	0	0	263
Promotes healthy child development	0	62	1040	0	0
Helps me promote healthy habits for children	0	0	1034	118	0
Makes me feel like I'm doing the best	250	0	855	400	273
Helps me live a healthier lifestyle	395	0	485	602	440
Makes me feel like a responsible parent	0	106	444	0	0
Is certified by a third party	0	179	0	1025	0
Receives positive ratings and reviews	69	146	12	649	136
Is recommended/endorsed by professionals	0	170	0	609	48
Offers the latest, most innovative technology	0	168	65	500	97
Has technology that is better for environment	0	180	207	418	120
Is a brand I've had a good experience with	0	115	115	0	474
Makes my water look clearer	333	162	0	0	445
Is a brand that I'm familiar with	0	116	81	0	411

SUMMARY OF SEGMENTS

We ended up extracting five segments of roughly equal size. The segments were intuitive and actionable given the client’s objectives.

- Basic Value Shopper (20%)—Important to be affordable and have low/cheap maintenance
- Turnkey Investor (22%)—Important to increase home value, save money, reduce work
- Family Guardian (20%)—Child development and family are important
- Informed Performance Seekers (17%)—Important to have recommended system with positive ratings and innovative technology
- Brand Validators (21%)—Important to be good brand and appealing design

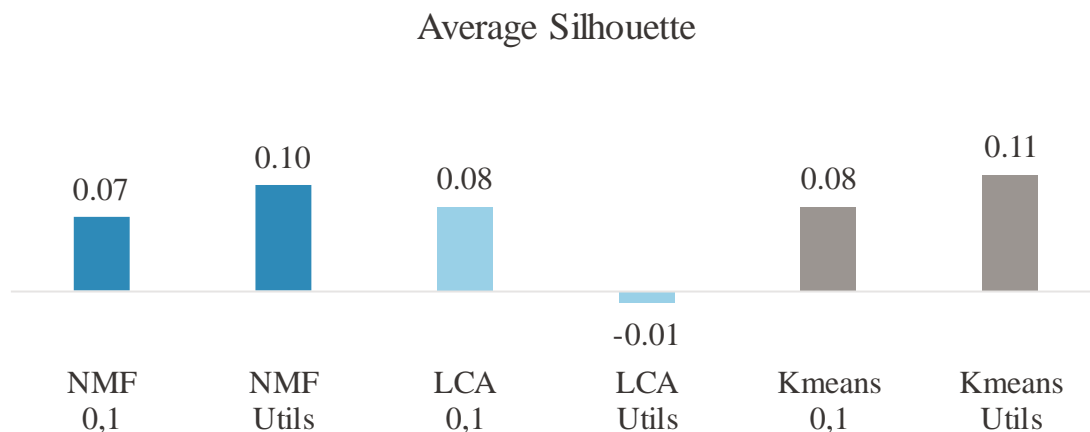
NMF VS. OTHER TECHNIQUES

We compared NMF to other approaches using both the binary recoded MaxDiff utilities and the original utilities. We analyzed cluster quality using a metric called a silhouette value, which looks at within-segment homogeneity versus between-segment heterogeneity.

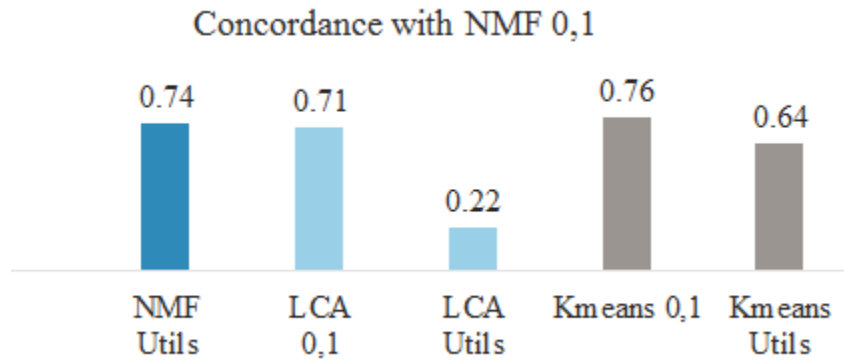
Silhouette values can range from -1 to 1 where the higher the value in the positive direction the better the quality of the solution. A “0” would indicate a respondent is just as similar to members of other clusters as they are to members within their cluster. A “1” would mean essentially everyone within the cluster is identical, and the clusters are completely different from one another (we know this is never really the case). Here we use silhouette values to compare the relative quality of the solutions.

$$sv = \frac{\text{separation} - \text{cohesion}}{\max(\text{separation}, \text{cohesion})}$$

Case Study Silhouette Values:



The solutions are similar in terms of cluster quality, with the exception of the Latent Class solution using original MaxDiff scores (which has no differentiation amongst the segments).



We also looked at the overlap, or concordance, of the segments derived via the various solutions to the segments that were extracted via NMF using binary data. For the most part there is a large overlap among the various solutions (64%–76% concordance rates), again with the exception of the Latent Class solution that uses the MaxDiff utilities (22% concordance rate). We also looked at the results via Hierarchical clustering and regardless of coding we consistently ended up with a single, large segment, so no additional analyses were run using the Hierarchical approach.

CONCLUSION

Non-negative Matrix Factorization has proven to be a very valuable segmentation technique across a number of disciplines and we believe that it offers distinct advantages in market research. A really nice benefit of NMF is not only do we see how the respondents group together but we also simultaneously see how the basis variables group. We have shown that in our simulated datasets, NMF performs just as well and often times better than the other segmentation techniques we looked at (K-means Clustering, Hierarchical Clustering, and Latent Class Mixture Model). In the case of our highly correlated datasets, NMF significantly outperforms the other methods. Like any other method, NMF comes with its limitations. For one, we are limited to using it only on non-negative datasets. R and Python are the only software packages we know that can currently do NMF. Nevertheless, we think NMF is an excellent technique to add to our segmentation toolkit.



Michael Patterson



Jackie Guthart



Curtis Frazier

REFERENCES

- Ketchen, David J., and Shook, Christopher, L. (1996), "The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique," *Strategic Management Journal*, 17:6 (June), 441–458.
- Lee, Daniel D., and Seung, H. Sebastian (1999), "Learning the parts of objects by non-negative matrix factorization," *Nature*, 401, 788–791.

VARIABLE SELECTION FOR MBC CROSS-PRICE EFFECTS

KATRIN DIPPOLD-TAUSENDPFUND

CHRISTIAN NEUERBURG

GfK

ABSTRACT

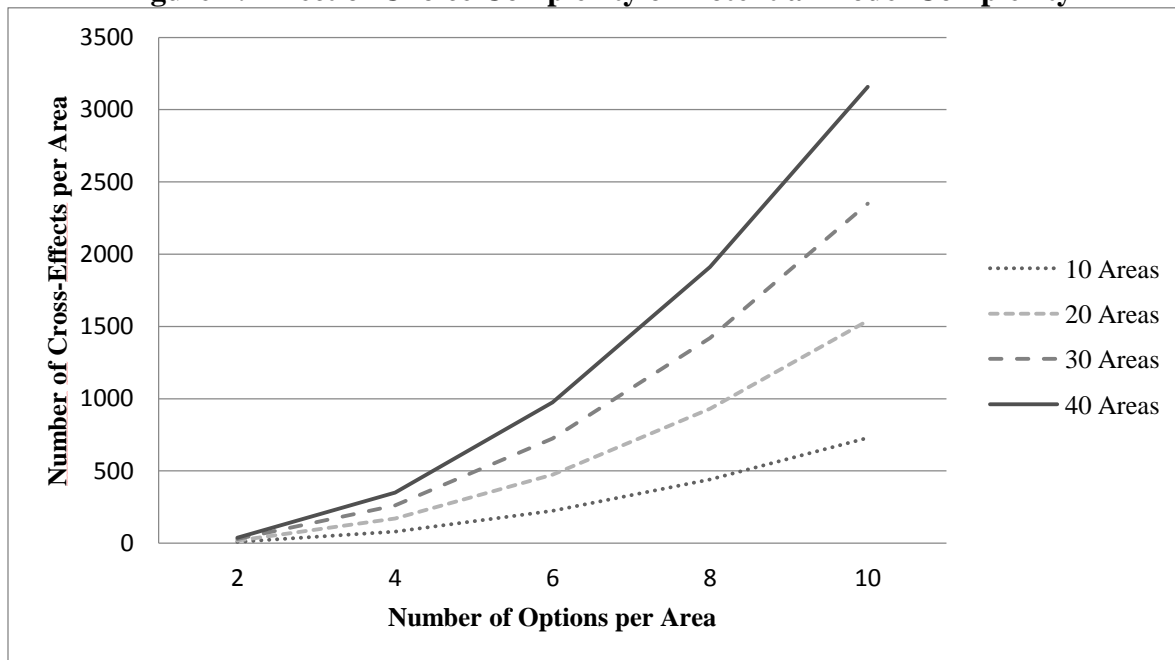
In Menu-Based Choice experiments, cross-price effects need to be selected carefully in order not to overfit the models or have simulation results distorted by “noisy” parameters. We investigate different approaches that support the selection of cross-price effects and compare their performance based on synthetic datasets under varying data conditions. We find that selection approaches that result in sparse models, e.g., variable selection with lasso, do very well under different data settings, especially with respect to our newly developed KPI that measures the quality of the resulting pricing decision. But also the relationship Chi² test, that is statistically less advanced and already implemented in the MBC software, performs very well if the p-value cut is selected carefully. We emphasize that complex choice menus require a strict variable selection.

MOTIVATION

Menu-Based Choice (MBC) allows to explore tasks where the choice can be described as a “pick-any”/“multi-check” situation comparable to the choice from a menu at a restaurant, where different menu areas are available (e.g., burgers and side orders) that consist of various priced options. Although MBC has become an established conjoint tool, the complexity of MBC models still makes high demands on the efficiency of estimation techniques and on an analyst’s modeling skills. The choice within a menu area is typically formulated as a multinomial logit (MNL) model. Different menu areas are in many cases linked via cross-price effects (“Serial Cross Effects”) to account for any effects between menu areas (Orme, 2010). These cross-price effects are defined in Sethuraman et al. (1999) as to “measure the effects of a brand’s price promotion (temporary price reduction) on a competitive brand’s market share.” The effects can either be substitutional or complementary in nature.

In this serial cross-effects setting, the number of potential cross-price effects that capture the interdependencies among offerings on a menu snowballs with the complexity of the tasks, i.e., the number of attributes (menu areas) and priced options (modular alternatives) in the design. This is because every priced option in a menu area can be potentially linked to every option in any other menu area. For a visual representation of this effect, please see Figure 1 (Neuerburg, 2013, p. 101). This visualization assumes that a none option is available in every menu area that does neither experience nor exert cross-price effects. This complexity prohibits estimating all possible cross effects in real-world scenarios. In addition, practical experience shows that “often 75% or more of potential cross effects will turn out to be non-significant” (Orme, 2012, p.33).

Figure 1: Effect of Choice Complexity on Potential Model Complexity



Leaving aside technical restrictions, estimation of a fully specified model is in most cases not desirable because of potential overfitting and a high level of noise that will affect simulation. Additionally, a simulation on a potentially misspecified model might also cause missed sales if price optimization was based on a wrong model.

Hypotheses on the existence of specific cross-price effects should ideally guide the selection of effects—but often these are not straightforward to derive, or clients want to follow a more data-driven approach. Thus, there is a strong need for efficient variable selection approaches prior to HB model estimation.

As a consequence, the problem of model pruning is approached with variable selection. Selection of cross-price effects with a χ^2 test is already implemented in Sawtooth Software's MBC tool. Additionally, we suggest three other approaches that can be fine-tuned to the needs of MBC choice modeling.

The primary objective of this research is to investigate systematically the performance of variable selection approaches applicable to data from MBC experiments. We also want to see how important model pruning is for the quality of models and simulations. We compare four different selection approaches and test them under varying data conditions. As a result, it can be derived if and under which conditions the tools available in Sawtooth Software's MBC software fail and have to be supplemented or substituted with a more sophisticated variable selection technique.

VARIABLE SELECTION APPROACHES

Variable selection comes into play prior to the HB estimation¹ of the MNL model so that only a pruned model specification enters the HB run. We test four different variable selection

¹ All findings are based on HB estimation, other estimation techniques, e.g., latent class or aggregate logit, are out of scope. We only use HB means not the draws themselves.

methods, three of them coming in two specifications each. The choice of methods was guided by the idea to come up with a selection toolbox that is quickly implemented for daily use and easily automatized for large-scale problems. All estimations were executed in R.

The first approach is the **relationship Chi² test** that is already employed in the Sawtooth MBC software and therefore can be considered the current industry standard. For every option in the menu and every potential cross-price effect, a crosstab is constructed from the data; the corresponding p value is calculated. As the MBC manual (Orme, 2012) suggests different p-value cuts, we explore—with values of 0.05 and 0.2—the extremes of the continuum to gain insights into how sensitive the selection performance reacts to the set cut values.

A very generic selection alternative is the estimation of an **aggregate logit model**. For this purpose, we decompose the described MNL models into binomial logit models with the choice of one priced option in a menu area as the binary dependent variable. Alternative-specific constant, own price effect and all possible cross-price effects enter the model as independent variables. The p-values of the cross-price effects are compared to a p-value cut to determine whether the effects are to be included into the model. Again, two different cuts are tested: 0.05 resulting in sparse models and 0.2 resulting in complex models.

The third variable selection method draws on ideas of Tibshirani (1996): a **Lasso** (least absolute shrinkage and selection operator) variable selection is performed with the R glmnet package (Hastie & Qian, 2014). Given a binomial logit model for every priced option in every menu area, glmnet allows to force the alternative-specific constant and own price effect into the model but selects which cross-price effect should be added into the model. The binomial logit models are fit via penalized maximum likelihood, which sets a selection of the model parameters to zero. The underlying idea of the penalty is that the sum of the absolute value of the model parameters may not exceed a given value. Strength of the overall penalty is fine-tuned with a parameter λ , which drives the model sparsity. The higher λ is, the sparser the resulting model we have. We test two specifications of λ building upon the λ values that glmnet generates as default:

- λ_{\min} , which is the λ value for the model with the smallest mean cross-validated error
- λ_{1se} , which gives an even sparser model within one standard error of the λ_{\min} model

We will use λ_{1se} resulting in an extremely sparse selection and the average of λ_{\min} and λ_{1se} (λ_2) resulting in a more generous selection as our tuning parameters.

The fourth selection approach to complement our selection toolbox is the R glmulti package (Calcagno & de Mazancourt, 2010). It is flexible and convenient to use, being a wrapper for glm. After several initial tests on various data sets, we found the genetic algorithm by a factor of approx. 15–20 slower than the short-cut alternative without delivering a superior parameter selection. Therefore, we applied the implemented short-cut method, an efficient **Branch-&-Bound algorithm** as last selection alternative. Identical to the Lasso estimation, binomial logit models were defined with the alternative-specific constant and own price effects forced into the model. The selection itself followed a branch-and-bound logic implemented in the leaps package.

As a side note, we refrain from variable selection with the likelihood ratio test. Building a model from scratch only referring to log likelihood ratios would be very time-consuming. Moreover, the outcome would potentially be affected by the order in which the cross effects were tested.

SYNTHETIC DATA SETS

We make use of synthetic MBC data to evaluate the performance of the different selection techniques under varying data settings. Most importantly, synthetic data allow us to know which cross-price effects really exist and which do not exist. So we can easily judge the quality of the selection for each technique comparing the selected cross-price effects to the real set of cross-price effects. The creation of the simulated data sets is largely based on the approach described in Neuerburg (2013). With regard to the cross-price effects we make the following simplifying assumptions:

- We randomly switch on two cross-price effects for each option within a certain menu area. A consequence is that the number of true cross-price effects per MNL model increases with menu complexity.
- In a second step, we randomly decide about substitutional or complementary relationships, i.e., a positive or a negative sign of the effect.
- Third, we make sure for each option that the own-price effect is potentially always stronger than the cross effects. We hypothesize that own price is—besides the alternative-specific constant—the main driver of choice.

The data conditions are varied in terms of complexity of the modeled menus, sample size, number of tasks and heterogeneity of the respondents. Complexity of the choice menu comes in three levels: 10 menu areas with 2 options each (1 priced option and 1 none), 15 menu areas with 3 options each (2 priced options and 1 none) and 10 menu areas with 6 options each (5 priced options and 1 none). The price levels for the different options in a menu overlap to a certain amount, but increase from option to option (see Table 1). This is, for example, the price definition for the most complex choice menu. The first line only would be the price definition for the least complex choice scenario. Please note that in our case, all price parameters have a linear formulation.

Table 1. Price Levels for Menu Areas

	Price Level 1	Price Level 2	Price Level 3	Price Level 4	Price Level 5
Option 1	5	8	10	12	15
Option 2	10	16	20	24	30
Option 3	15	24	30	36	45
Option 4	20	32	40	48	60
Option 5	25	40	50	60	75

We test four different sample sizes: 100, 250, 500 and 1000. We present the results for either 5 tasks per respondent or 10 tasks per respondent. Finally, the cross-price structure can be either homogeneous or heterogeneous. In the first case, cross-price effects are identical for all respondents (same effects, same sign). In the second case, a respondent can belong to one of three segments with a specific cross-price structure. Segment sizes are 45%, 35% and 20%. This leaves us with $3 \cdot 4 \cdot 2 \cdot 2 = 48$ different data settings. A summary of the data settings can be found in Table 2.

Table 2. Dimensions of Synthetic Data

Data Dimension	and its levels			
Sample Size	100	250	500	1000
Number of Tasks	5 tasks		10 tasks	
Heterogeneity	Homogeneous CP effects		Heterogeneous CP effects	
Complexity	10 areas with 2 items	15 areas with 3 items	10 areas with 6 items	

SELECTION KPIS

The performance of the different variable selection techniques is evaluated based on all synthetic datasets. This allows us to compare the identified cross-price effects directly to the synthetic “true” cross-price effects. Four outcomes of the comparison are possible (Table 3):

Table 3. Development of KPIS

Cross-price effect ...		SYNTHETIC TRUTH	
		... existing	... not existing
MODEL-DRIVEN SELECTION	... selected	TRUE POSITIVE	FALSE POSITIVE
	... not selected	FALSE NEGATIVE	TRUE NEGATIVE
		SENSITIVITY	SPECIFICITY

Based on this comparison, these two measures are chosen as key performance indicators for the goodness of the selection:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Our first KPI **sensitivity** focuses on whether existing effects are indeed discovered and is therefore the share of all the existing effects selected as existing. The second KPI **specificity** attaches importance to de-selecting non-existing effects from the model and is formulated as the portion of all the non-existing effects estimated as non-existing. As both KPIS are shares, they range between 0% and 100%. In addition, the number of selected cross effects is reported as an indicator of model sparsity or complexity (see Table 4):

Table 4. Performance of Selection Techniques on KPIs

	# CPs	Sensitivity	Specificity
Chi²-Test (p-value = 0.2)		43%	90%
Chi²-Test (p-value = 0.05)		28%	98%
Aggregate Logit (p-value = 0.20)		56%	86%
Aggregate Logit (p-value = 0.05)		40%	97%
LASSO glmnet (lambda = $l_{1.se}$)		3%	100%
LASSO/glmnet (l_2 = average of l_{min} and $l_{1.se}$)		35%	98%
Branch&Bound/glmulti		52%	90%

Key

	Below the real number of CP effects
	Close to the real number of CP effects
	Above the real number of CP effects

Looking at the number of cross-price effects per menu area, we find that only the strictest Lasso specification with λ_{1se} selects fewer than the real number of effects. The Chi² test with a p-value cut of .05 and the more generous Lasso formulation with λ_2 come closest to the real number of cross-price effects. All other methods select more than the real number of cross-price effects into the model, aggregate logit with a p-value cut of .2 resulting in by far the most effects.

The results for sensitivity and specificity are in line with what we see for the number of cross-price effects. For sensitivity, we have a clear winner: aggregate logit with a p-value cut of .2, which we already have seen delivers the most complex models, results in the highest sensitivity value followed by the branch & bound algorithm. The strict Lasso specification that conveys the sparsest models nearly neglecting all cross-price effects clearly performs worst and does not show any sensitivity. The specificity results are the sensitivity results vice-versa. The strict Lasso selection nearly without cross-price effects reaches the highest specificity value possible. By far the worst specificity value follows from the aggregate logit model de-selecting effects only with a .2 threshold.

These results indicate a trade-off between sensitivity and specificity. We do not have a selection approach that can perform very well for both KPIs. But if we have to trade-off sensitivity against specificity, how do we know what is more important for a “good” variable selection approach? Is it sensitivity, which means finding as many existing effects as possible but accepting at the same time to estimate many non-existing effects as well? This will lead to very

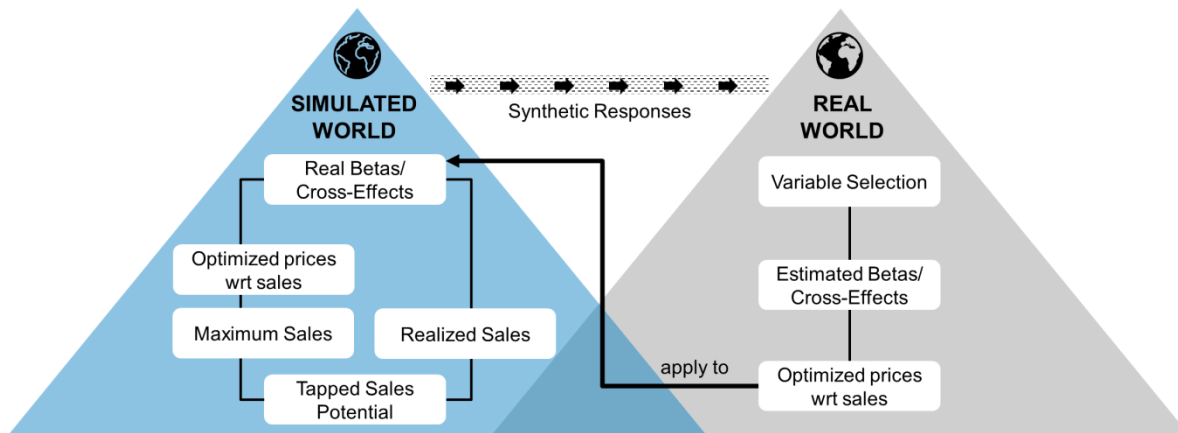
complex models, which take a long time to estimate, may be counter-intuitive in the interpretation and are—most importantly—eventually noisy in the simulation? Or should we stress simplicity of the models and favor specificity? That would mean that we might possibly neglect existing effects but come up with a sparse, less noisy model in the end.

Our selection KPIs do not help us to answer this question. As a kind of “reality check,” we develop a new KPI that measures the quality of the managerial decisions that are based upon our suggested models. In this case, the managerial decision to be made from the MBC models is the pricing of the choice menu.

TAPPED SALES POTENTIAL KPI

We develop a KPI that compares the sales that arise from the prices optimized for the selection models to the maximum achievable sales that arise from the prices optimized for the true market response. In other words, the new KPI measures what portion of the sales can be tapped—in relation to the potential maximum—by offering the choice menu at the prices that were suggested by the models defined with the different variable selection approaches. This KPI is calculated for every selection approach and every data setting.

Figure 2. KPI Development



The development of the KPI follows these steps (flow in Figure 2):

- Variable selection is performed for a synthetic data set following a specific selection algorithm. The specified models are estimated with a hierarchical Bayes routine. The prices are then optimized with respect to a sales maximum of a menu while setting the min and max prices as lower and upper bounds (see right triangle in chart).
- Secondly, the prices are set to maximize sales for the synthetic/known models. The result is a sales figure, called “maximum sales.” It will be the benchmark that the sales based on the estimated models will be compared against (see left triangle, left-hand side).
- Actually occurring sales for the prices from the estimated models are determined by plugging the prices optimized for the estimated sales function to the synthetic/known sales function, as this is the true market response. This will deliver a result, we call “realized sales” (see left triangle, right-hand side).

- The new “Tapped Sales Potential” KPI is then “realized sales” divided by the “maximum sales” (see left triangle, bottom).

This percentage will then tell us how good the models derived from our different selection methods are in coming close to our synthetic reality. Finally, we will be able to judge whether sparse models with high specificity or complex models with high sensitivity do a better job. Furthermore, the KPI is calculated for a baseline model, the sparsest model possible that completely omits any cross-price effects.

RESULTS

The percentage values shown in Table 5 are the “tapped sales potential” KPI, i.e., how much of the maximum sales can be tapped with the models of the specific selection method. Values in line 2 are shown as mean across all 48 data scenarios (i.e., 3 complexity levels, 2 heterogeneity levels, 4 different sample sizes, 2 different task numbers). In the lines below, the minimum and maximum values that were achieved by the method in a specific data scenario are given. The respective winner is marked in bold. What is striking at first glance, is that the mean values differ but the range is quite narrow. Averaging over all data scenarios, we see that the sparsest models perform best, i.e., the strict lasso, the Chi² test and the baseline model. Methods that result in more complex models, e.g., branch & bound and aggregate logit with cut-off at .2, are clearly the losers.

Table 5. Tapped Sales Potential KPI by Methodology

	Baseline	Chi ² (0.20)	Chi ² (0.05)	AggLog (0.20)	AggLog (0.05)	Lasso (λ_{1se})	Lasso (λ_2)	B&B
Average (48 scen)	93%	93%	93%	89%	91%	93%	92%	89%
Minimum (48 scen)	85%	80%	82%	78%	81%	83%	82%	77%
Maximum (48 scen)	97%	98%	98%	96%	96%	97%	97%	96%

The real differences, however, become apparent, if looking at the minimum and maximum values occurring across the 48 different data scenarios. Not only the ranges differ from selection approach to selection approach but also the data setting itself seems to have an impact on the performance of a selection technique.

In order to investigate these effects further, we look into specific data scenarios (

		Baseline	Chi ² (0.20)	Chi ² (0.05)	AggLog (0.20)	AggLog (0.05)	Lasso (λ_{1se})	Lasso (λ_2)	B&B
Low complexity	Hom	88%	95%	95%	96%	95%	91%	95%	96%
	Het	87%	90%	88%	90%	91%	87%	89%	89%
High complexity	Hom	95%	96%	96%	90%	93%	95%	97%	93%
	Het	97%	97%	97%	88%	95%	97%	95%	93%

). First, we fix sample size to 500 and number of tasks per respondent to 10. These settings should avoid results becoming distorted by too little sample or too few answers per respondent. An exploration of all results showed choice complexity and heterogeneity of cross-price effects to be the main drivers of the “tapped sales potential” KPI. Therefore, these data dimensions are investigated in detail. The complexity of the choice menu is varied from low (10 menu areas with 2 options) to high (10 menu areas with 6 options). Additionally, we differentiate the level of heterogeneity in the cross-price structure (homogeneous vs. heterogeneous).

Table 6. KPI by Scenario and Methodology

		Baseline	Chi ² (0.20)	Chi ² (0.05)	AggLog (0.20)	AggLog (0.05)	Lasso (λ_{1se})	Lasso (λ_2)	B&B
Low complexity	Hom	88%	95%	95%	96%	95%	91%	95%	96%
	Het	87%	90%	88%	90%	91%	87%	89%	89%
High complexity	Hom	95%	96%	96%	90%	93%	95%	97%	93%
	Het	97%	97%	97%	88%	95%	97%	95%	93%

Scenario 1. In the low complexity scenario with homogeneous cross-price structure, the “high sensitivity” methods returning more complex models, e.g., branch & bound or aggregate logit with p-value cut at .2, perform best. The baseline-model, which is the winner when averaging across all data scenarios, performs worst. The KPI values range is very broad (88%–96%). We reason that in this simple data setting neglecting cross-price effects diminishes the quality of the pricing extremely.

Scenario 2. The second scenario is identical to scenario 1 with the cross-price effects structure now being specified for three sub-samples of respondents. All methods suffer from the increase in heterogeneity. Especially, branch & bound is no longer the top approach; aggregate logit and Chi² test with a more generous p-value cut do best. Compared to scenario 1, the range of KPI values becomes smaller (87%–91%).

Scenario 3. Switching to a scenario with high choice complexity but homogeneous cross-price structure, Lasso with λ_2 and Chi² do best. We also see that the baseline model without cross-price effects does quite well. The clearly worst performance comes from branch & bound and aggregate logit, especially with the loose p-value cut of .2.

Scenario 4. Here the methods resulting in sparse models, i.e., the strict lasso and especially both Chi² tests do very well. They seem to be the methods that make the most of the complicated underlying data structure. Especially the results of Chi² tests are not biased by the heterogeneity of the cross-price effects. The very good performance of the baseline model indicates that—in case of question—it might be better not to estimate any cross-price effects than to include wrong ones into the model, i.e., a clear point for model specificity. Whether omitting cross-price effects is practical in reality, where revealing cross-price effects is exactly what the users of simulators want to see, can be doubted. Interestingly, we do not see the same pattern as in the low complexity scenario, when the increase in heterogeneity automatically leads to a decrease in the KPI. Comparing performance of the selection approaches from scenario 3 to scenario 4, we rather see quite similar magnitudes.

CONCLUSIONS

Looking over all 48 data scenarios, methods that deliver sparse models, e.g., Lasso with a strict lambda penalty or Chi² test with a strict p value cut (0.05), perform best with regard to the “tapped sales potential” KPI. This finding gains importance as the choice menu increases in complexity. Therefore, for the HB models we tested, we conclude that we should give specificity a higher weight than sensitivity (favoring more parsimonious models to avoid overfitting) if choosing a selection approach.

Looking at specific scenarios, we have seen that all methods have data settings where they perform strongly or poorly. It is not easy to promote one method as a kind of “one fits all.” We definitely see that complex choice menus demand a very strict variable selection. That is why the

researcher should adapt his choice of selection approach to the complexity level of the choice task.

Model pruning clearly improves the quality of process and results. We see that sparse models, sometimes even the null model, perform very well with respect to our “tapped sales potential” KPI. Besides this, sparse models are easier to estimate, to interpret, and less noisy in the simulation.

We do not necessarily see a clear need for more sophisticated variable selection techniques. The χ^2 test performs well to very well, especially compared to methods that are statistically far more advanced and far more challenging in the implementation. Besides, the χ^2 test works under a multitude of data scenarios. This is of high importance, as many data dimensions, e.g., heterogeneity of the sample, might not be under the researcher’s control or not even be directly observable. We conclude that the χ^2 test is definitely a safe choice for selecting cross-price effects when the p-value cut is set carefully.

LIMITATIONS AND FUTURE RESEARCH

Limitations

In our research, we rely on synthetic data. Therefore, we cannot exclude that we might see different interactions of data dimensions and selection techniques when working with data from real respondents, for instance with respect to the number of tasks (“fatigue effect”). Besides, our results rely heavily on the assumptions we made for the cross-price effects structure, especially how sparse we assume it to be. For a data setting with a less sparse cross-price effect pattern, selection techniques that favor sensitivity might bring advantages. Finally, our findings are limited to HB estimation of the utilities. Other estimation techniques than HB might lead us to different conclusions. For instance, estimating the MBC models with aggregate logit might pose less strict demands on variable selection. We also work only with HB point estimates and do not make use of the broader information contained in the HB draws. We assume, however, that this disadvantages all tested methodologies in the same way, so we should not see a different ranking of methods if using draws instead of point estimates.

Future Research

A possible direction is to investigate other variable selection techniques. First, genetic algorithms are still worthwhile to explore more deeply. Other genetic algorithms than the `glmulti` package might be tested, or the `glmulti` genetic algorithm might work well for other data settings. Second, Bayesian variable selection that is already successfully applied in Market Basket Analysis, a closely related stream of research, might boost the quality of selection (Dippold & Hruschka, 2013). Finally, the most fundamental alteration in research would be to determine interdependencies in choices directly without the detour over prices: the modelling structure might be changed from serial effects models for each menu area to a joint model for all menu areas. In this way, interdependencies could be derived from the choice information directly. Especially the auto-logistic model comes with a solid methodological reasoning, e.g., Kamakura & Kwak (2015) or Kosyakova et al. (2017).



Katrin
Dippold-Tausendpfund



Christian Neuerburg

REFERENCES

- Calcagno, Vincent, and Claire de Mazancourt (2010), “glmulti: an R package for easy automated model selection with (generalized) linear models.” *Journal of Statistical Software*, 34(12), 1–29.
- Dippold, Katrin, and Harald Hruschka (2013), “Variable selection for market basket analysis.” *Computational Statistics*, 28(2), 519–539.
- Hastie, Trevor, and Junyang Qian (2014), *Glmnet Vignette*, Stanford, [available at https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html].
- Kamakura, Wagner A. and Kwak, Kyuseop (April 15, 2012), “Menu-Choice Modeling.” [available at SSRN: <https://ssrn.com/abstract=2162019> or <http://dx.doi.org/10.2139/ssrn.2162019>].
- Kosyakova, Tetyana, Thomas Otter, Sanjog Misra and Christian Neuerburg (April 23, 2017), “Measuring Substitution and Complementarity Among Offers in Menu Based Choice Experiments,” [available at SSRN: <https://ssrn.com/abstract=2957105> or <http://dx.doi.org/10.2139/ssrn.2957105>].
- Neuerburg, Christian (2013), *Modellierung von Wahlverhalten in modularen Auswahl-situationen. Ein simulationsbasierter Vergleich verschiedener Modellvarianten unter Berücksichtigung der Zahlungsbereitschaft*. Nuremberg: GfK-Verein.
- Orme, Bryan K. (2010), “Menu-Based Choice Modeling Using Traditional Tools,” in *Proceedings of the Sawtooth Software Conference 2010*, Sawtooth Software Inc, ed. Sequim, WA, 37–57.
- (2012), “Menu-Based Choice (MBC) for Multi-Check Choice Experiments,” (accessed May 1, 2015), [available at <http://www.sawtoothsoftware.com/download/mbcbooklet.pdf>].
- Sethuraman, Raj, Vivek Srinivasan, and Doyle Kim (1999), “Asymmetric and neighborhood cross-price effects: Some empirical generalizations,” *Marketing Science*, 18(1), 23–41.
- Tibshirani, Robert (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

ACCOMMODATING MULTIPLE DATA PATHOLOGIES IN CONJOINT STUDIES VIA CLEVER RANDOMIZATION AND ENSEMBLING

JEFFREY P. DOTSON

BRIGHAM YOUNG UNIVERSITY

ROGER A. BAILEY

THE OHIO STATE UNIVERSITY

MARC R. DOTSON

BRIGHAM YOUNG UNIVERSITY

1. INTRODUCTION

Ensemble-based approaches currently dominate the world of competitive out-of-sample prediction. From Kaggle to the Netflix Prize, the predictive power inherent in using many models overshadows prediction reliant on the performance of a single model. The primary reason ensembles predict so well is that they serve as a hedge against model misspecification. Since we have uncertainty about the correct model for any given context, running many models and producing a consensus is a simple yet powerful way to improve predictions.

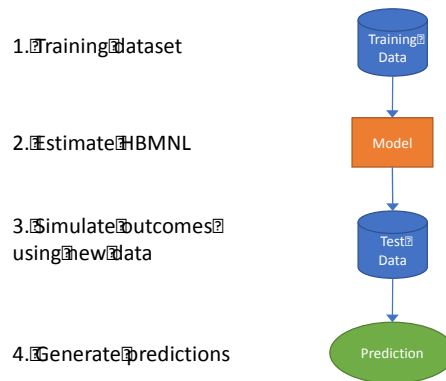
In the world of conjoint, most studies are conducted using a single model. When the aim of a conjoint study is solely inference and not prediction, a single-model approach is arguably best. The academic literature for conjoint is filled with models designed to improve inference, especially when respondents behave in ways that are “pathological” to the standard model. However, there are three reasons to argue for an ensemble-based approach to conjoint analysis. First, the end goal of many conjoint studies is prediction in the form of accurate market simulations. Second, we still have uncertainty about the correct model for any given conjoint study. Third, there is no single model that accounts for all the respondent behaviors that result in the “data pathologies” that have been addressed separately in the literature.

The remainder of the paper will be organized as follows. In Section 2, we walk through ensemble approaches to prediction. In Section 3, we detail our ensemble approach to conjoint analysis. In Section 4, we provide results from simulation studies and an empirical application. In Section 5, we conclude.

2. ENSEMBLE APPROACHES TO PREDICTION

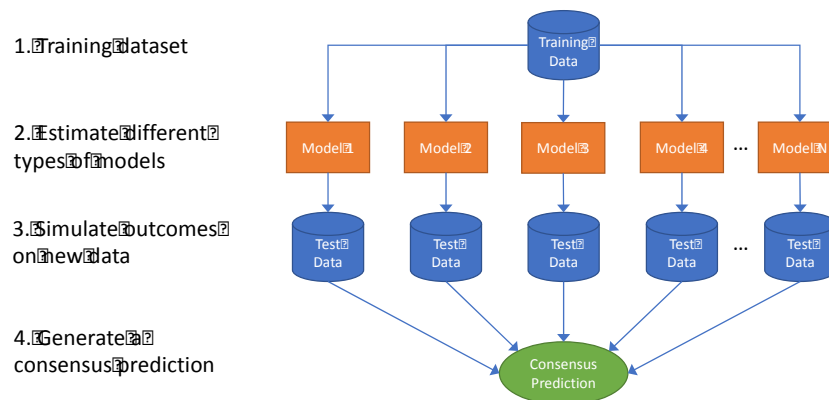
Before walking through ensemble approaches to prediction, it’s helpful to review the terminology commonly used in this space. The single-model approach to prediction is illustrated in Figure 1. The steps are to, first, specify the data used to train (i.e., estimate) the model; second, train the model; third, simulate outcomes using parameter estimates and test data (e.g., holdout tasks or holdout respondents); and fourth, use these simulated outcomes along with the test data to compute a prediction (e.g., hit rates).

Figure 1. A single-model approach to prediction.



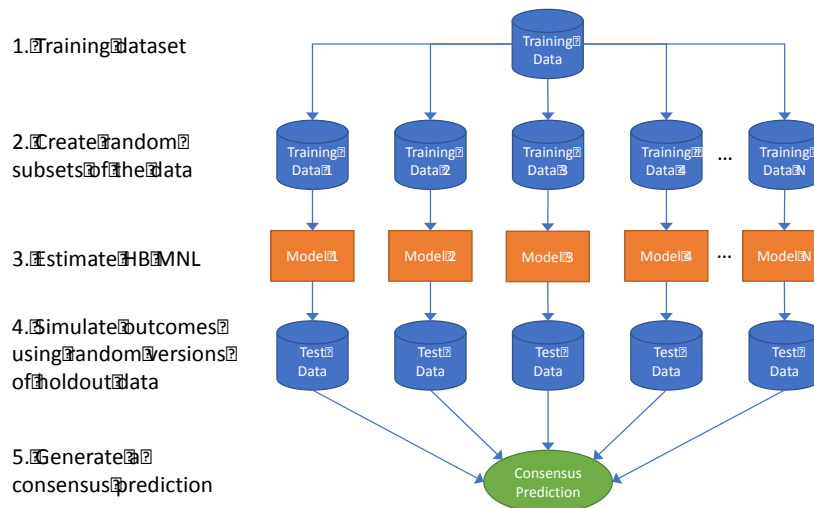
The single-model approach to prediction is the standard for conjoint studies. One notable exception is Kevin Lattery’s Sawtooth Software Conference 2015 presentation and paper, “A Machine Learning Approach to Conjoint.” In that paper, Lattery implemented an ensemble approach to prediction for conjoint analysis, which is illustrated in Figure 2. Here we can see that many models are fit and many predictions are calculated, each model using its own randomly selected test data. Finally, a consensus prediction is formed via aggregating the separate predictions (e.g., averaging predictions or taking the modal prediction).

Figure 2. An ensemble approach to prediction.



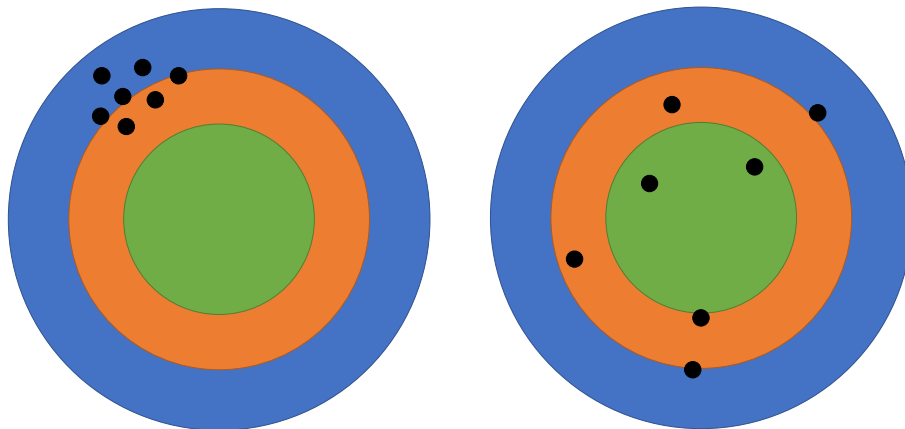
Lattery finds substantial improvements in prediction as the number of models in the ensemble increases. However, the ensemble isn’t theoretically grounded. Furthermore, the approach to prediction is fairly non-standard. A more general ensemble approach to prediction is illustrated in Figure 3. The additional step is to create a random subset of the data to serve as the training data for each of the models in the ensemble. Having separate training and test data for each of the models in the ensemble is the standard approach to ensembles.

Figure 3. An alternative ensemble approach to prediction.



Even with a non-standard ensemble approach to prediction and no theoretical justification for the ensemble, Lattery still found improvement in out-of-sample prediction. As stated previously, ensemble approaches to prediction are powerful because they serve as a hedge against model misspecification. This is often justified by ensembles striking an optimal balance on the bias/variance frontier. For example, in Figure 4 we can see two hypothetical targets. On the left we have low variance, high bias performance that represents using a single albeit misspecified model for prediction. On the right we have a high variance, low bias performance that represents the ensemble approach to prediction. The ideal is to have low bias and minimal variance. Our aim is to approach this ideal by using an ensemble that is theoretically justified.

Figure 4. Bias variance trade-off in ensembles.



3. ACCOMMODATING DATA PATHOLOGIES

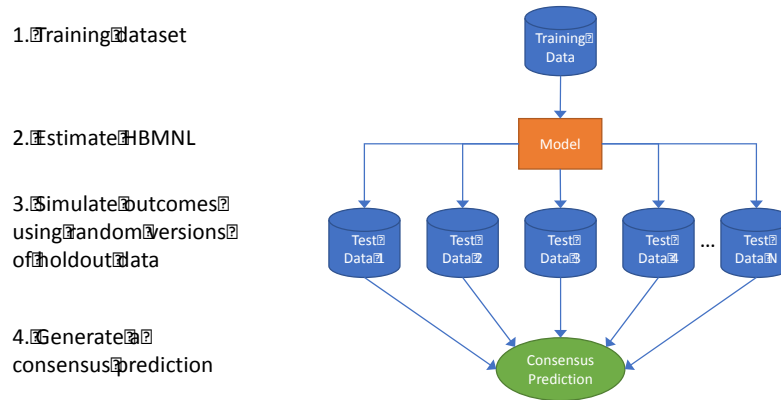
Our ensemble-based approach to prediction, like Lattery's, is non-standard. However, our approach differs in that we build an ensemble that is theoretically justified. In particular, we introduce randomization in our ensemble that is "clever" insofar as the randomization accounts for a potential data pathology. As noted above, much work has been done to build models that separately account for respondents producing data that is pathological to the standard model (i.e., can't be accounted for by the standard model and thus impedes prediction). Our use of an ensemble and clever randomization allows us to accommodate multiple data pathologies.

The standard model is a lower-level random utility model with an upper-level model over preferences. Consumers are able to assess the "utility" of each alternative in a choice set and pick the alternative that provides the greatest level of utility. Utility itself is made up of two components: A deterministic component and a random (to the researcher) component, where the deterministic component is expressed as a (linear, compensatory) function of the design of the alternative and the random component is assumed to come from an independent and identically distributed Gumbel distribution. Finally, while we estimate preferences at the individual level, we assume that the preferences of all individuals are drawn from a common multivariate normal distribution.

Development in the academic conjoint literature has focused on addressing specific data pathologies separately (e.g., attribute non-attendance, screening rules, poor respondent quality, non-IIA choice behavior, respondent fatigue, and alternative decision rules). These models fit the data better and provide marginal improvements in prediction. Although prospectively useful, especially in terms of inference, these models are rarely used in practice for three reasons. First, they are theoretically and computationally complex (i.e., difficult to understand and time-consuming to estimate and simulate). Second, we don't have high-quality commercial software that can be used to fit these models. Third, each model deals with a single data pathology. The challenge of model misspecification persists wherein, a priori, it is hard to know which pathology will prove problematic (i.e., which model should be fit). This problem is further complicated if, as might be expected, multiple pathologies are present in a single dataset.

In our ensemble-based approach to prediction, we don't have to fit complicated models, but we do need to compute a lot of predictions! The trade-off is between model complexity and computational intensity. Figure 5 illustrates our approach. Note that this is non-standard, in that we have both a single training dataset and a single model. The randomization we introduce is at the level of the randomly selected test data. We will accommodate two data pathologies, attribute non-attendance and screening behavior, separately and jointly.

Figure 5. Our ensemble approach.



Pathology 1: Attribute Non-Attendance

Attribute non-attendance is when respondents ignore subsets of attributes when making decisions (i.e., part-worths are 0 for all levels of the attribute). To accommodate this data pathology, we randomly set the part-worths for all levels of an attribute to 0 across test datasets. To be clear, we implement the following:

1. Estimate an HB MNL on training data
2. Loop over respondent-level part-worth estimates
 - randomly select an attribute
 - with a given probability, set all part-worth estimates for that attribute to 0
3. Predict first choices (e.g., max utility) for each choice set in the test data
4. Repeat steps 2 and 3 many times
5. Generate a consensus (e.g., most commonly selected) prediction

Pathology 2: Screening Behavior

Screening behavior is when respondents use certain attribute levels to screen out alternatives from consideration (i.e., part-worths are approximately negative infinity for all levels being screened on). To accommodate this data pathology, we randomly set the part-worths for levels to approximately negative infinity across test datasets. To be clear, we implement the following:

1. Estimate an HB MNL on training data
2. Loop over respondent-level part-worth estimates
 - randomly select an attribute level
 - with a given probability, set the part-worth estimate for that level to approximately negative infinity
3. Predict first choices (e.g., max utility) for each choice set in the test data
4. Repeat steps 2 and 3 many times
5. Generate a consensus (e.g., most commonly selected) prediction

Joint Ensemble for Pathologies 1 and 2

To accommodate for both data pathologies, we implement the following:

1. Estimate an HB MNL on training data
2. Loop over respondent-level part-worth estimates
 - randomly select an attribute level
 - with a given probability, set the part-worth estimate for that level to approximately negative infinity
 - randomly select an attribute
 - with a given probability, set all part-worth estimates for that attribute to 0
3. Predict first choices (e.g., max utility) for each choice set in the test data
4. Repeat steps 2 and 3 many times
5. Generate a consensus (e.g., most commonly selected) prediction

4. SIMULATION STUDIES AND EMPIRICAL APPLICATION

Four simulation studies demonstrate the potential of our ensemble approach to prediction. Figure 6 shows that when neither of the two data pathologies are present in the simulated data, the lower-level model (i.e., the standard model) and our ensemble approach predict about the same. Figure 7 shows that when attribute non-attendance is present but screening is not, our ensemble approach slightly out predicts the standard model. This is repeated with Figure 8 when screening is present but attribute non-attendance is not. However, Figure 9 clearly demonstrates the benefit of the approach as we see a large jump in predictive ability for our ensemble approach when both of the data pathologies are present in the simulated data.

Figure 6. Simulated Data: No Attribute Non-Attendance + No Screening.

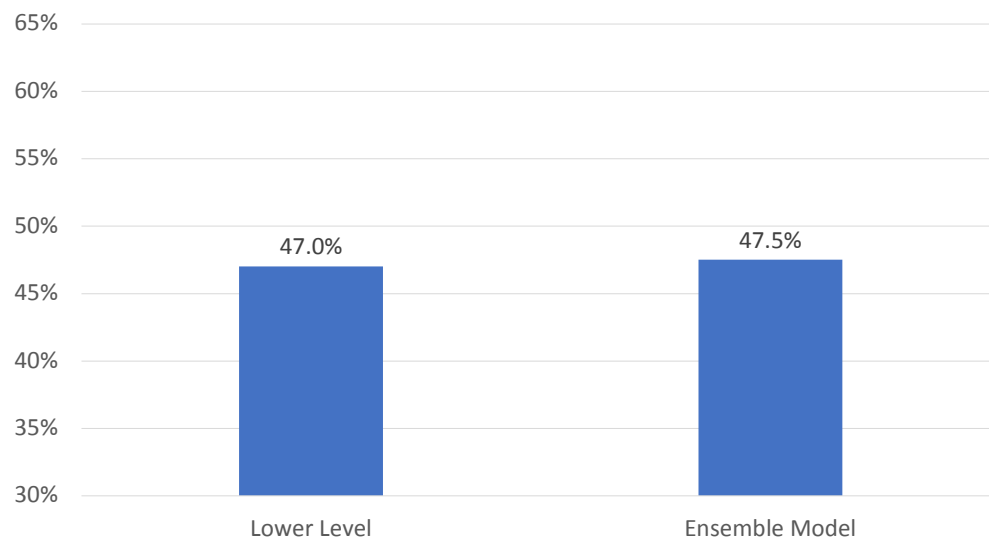


Figure 7. Simulated Data: Attribute Non-Attendance without Screening.

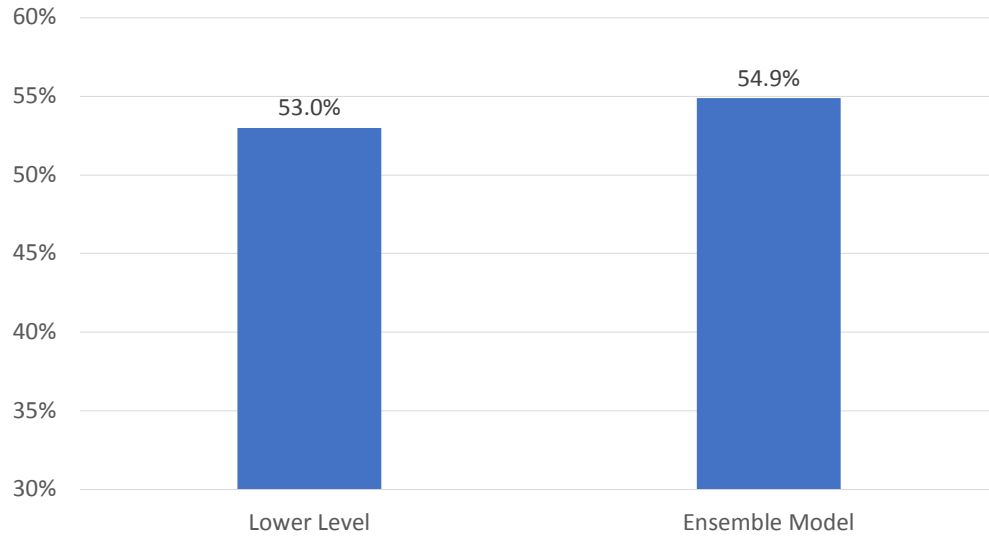


Figure 8. Simulated Data: Screening without Attribute Non-Attendance.

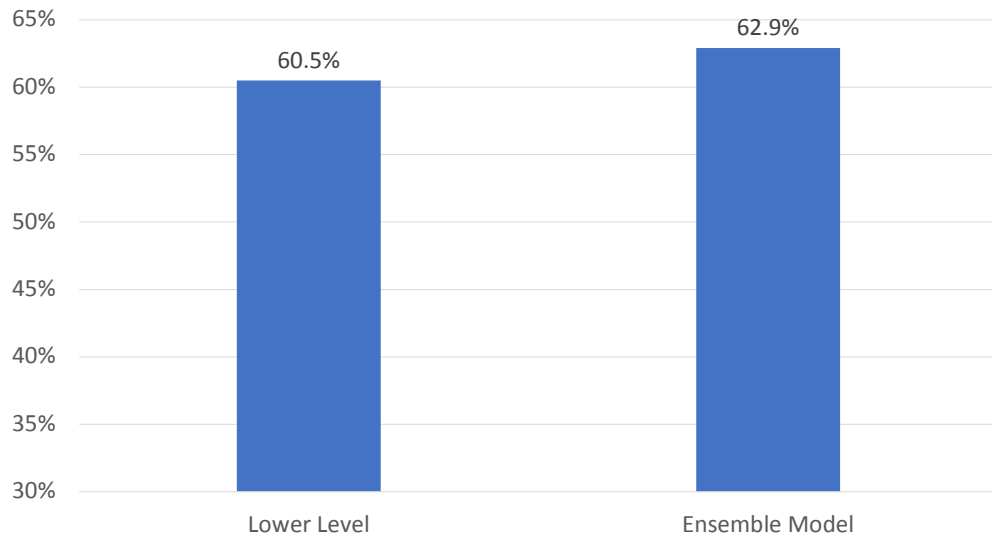


Figure 9. Simulated Data: Attribute Non-Attendance + Screening.

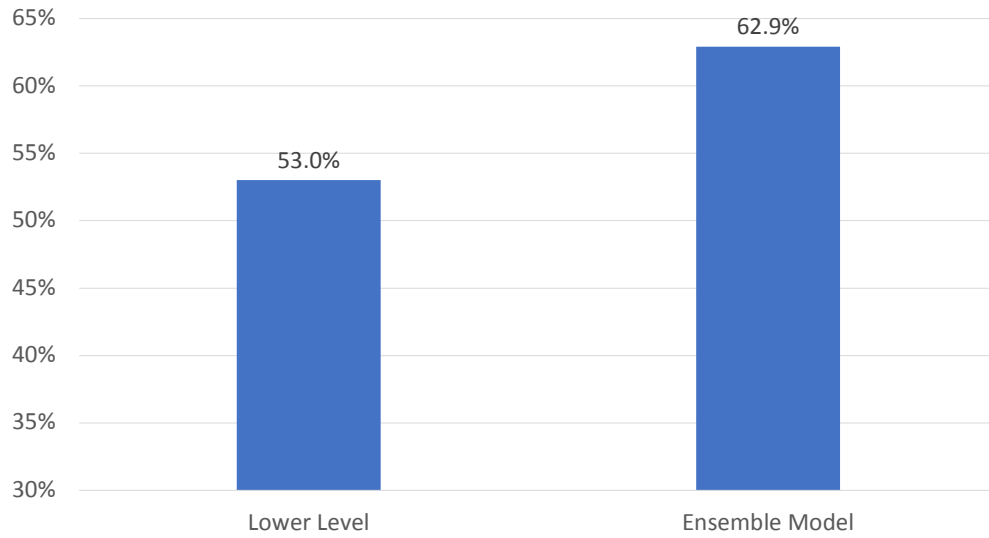
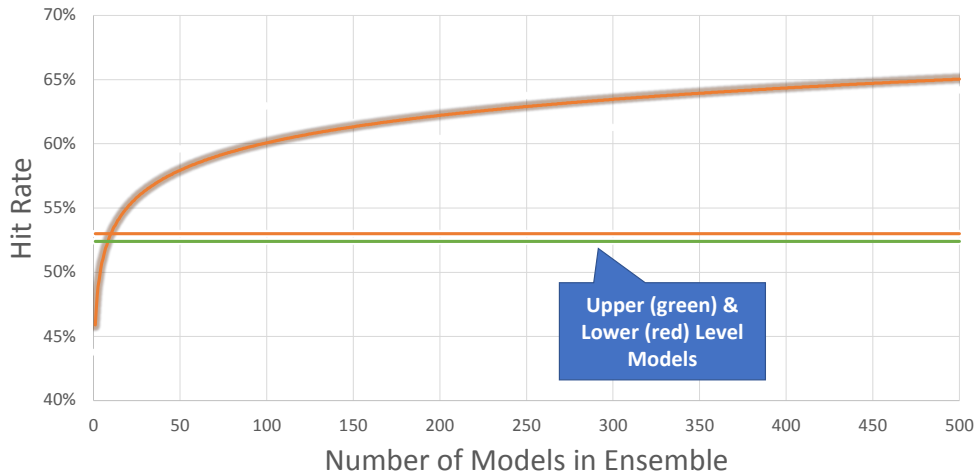


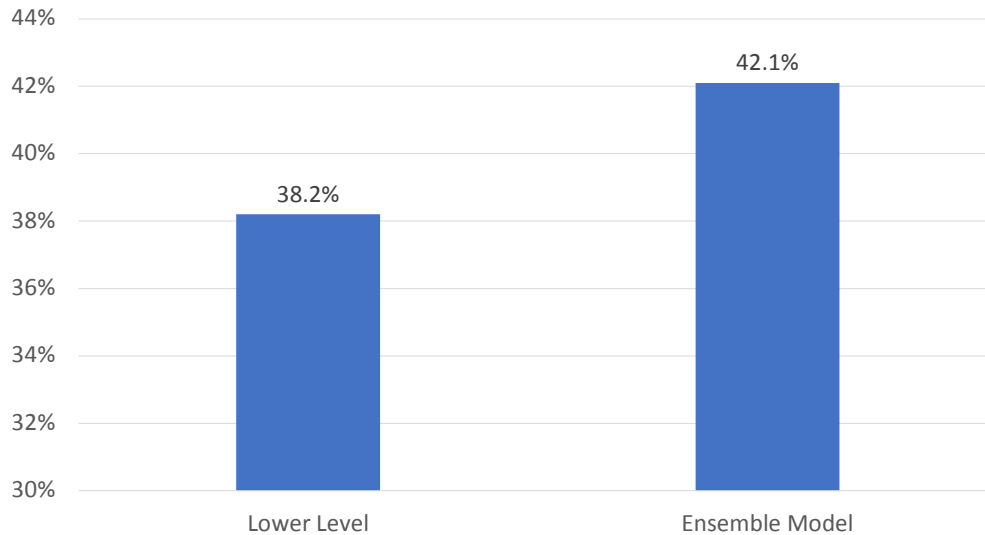
Figure 10 shows that the improvement in prediction increases as more test datasets following the theoretical justification outlined above are added to the ensemble.

Figure 10. Predictive fit as a function of ensemble size.



Finally, Figure 11 shows that the improvement in prediction for our ensemble approach over the standard model is clearly present for real data. Interpolating from our simulation experiments, this is most likely attributed to both of the data pathologies being present in the data, a condition that isn't accounted for by the more attribute non-attendance and screening models separately.

Figure 11. Performance on actual data.



5. CONCLUSION

When the goal for a conjoint study is out-of-sample prediction, there is great potential in an ensemble approach. The benefit of our ensemble approach is its construction is theoretically justified, simple to implement, and performs especially well when multiple data pathologies are present in a dataset.

There are a variety of next steps to consider. Our stylized, non-standard approach where clever randomization is induced only for the test data should be expanded to allow for randomization of training data and the training of many models. This will necessitate faster or more efficient computation. Given the benefit we've seen with accommodating only two data pathologies, more pathologies need to be considered and accounted for. Finally, introducing smarter ensembles and better prediction aggregation can only improve the approach, especially if the method of aggregation allows us to retain the benefits of inference and not simply produce improved predictions.



Jeffrey P. Dotson



Roger A. Bailey



Marc R. Dotson

TOOLS FOR DEALING WITH CORRELATED ALTERNATIVES

KEVIN LATTERY
JEROEN HARDON
SKIM GROUP

1.0 BACKGROUND: THE RED-BUS/BLUE-BUS PROBLEM

Most practitioners are familiar with the so-called “Red-Bus/Blue-Bus Problem.” The underlying property leading to this problem is termed “Independence from Irrelevant Alternatives” (IIA). The basic idea of IIA is that the ratio of any two products’ shares should be independent of all other products. This sounds like a good thing, and at first, IIA was regarded as a beneficial property.

However, if we look at another way, we see that an improved product gains share from all other products in proportion to their original shares. When a product loses share, it loses to others in proportion to their original shares. Stated that way, it is easy to see that IIA implies an unrealistically simple model in terms of sourcing. In the real world, products compete unequally with one another and when an existing product is improved, it usually gains most from a subset (nest) of products with which it competes most directly.

After the introduction of hierarchical Bayes we could estimate respondent-level betas. This moves IIA to the respondent level, but significantly reduces IIA at the overall level. In many cases, this may be sufficient for sourcing. In some cases however respondent-level IIA is still too strong of an assumption. Here are some examples we have observed when respondent-level IIA still leads to problems:

- Adding 30 versions of the same product in your simulator and it can (and most likely will) dominate the market.
- Related to the above, when doing a portfolio optimization, adding similar products will result in overly inflated market share.
- When we have strong category differences and we expect there is little to no sourcing between them.
- When we test lots of SKUs in a category.

In this paper we are comparing different methods that could potentially (partly) solve the “Red-Bus/Blue-Bus Problem.” We tested the following methods:

1. Standard HB
2. Post hoc Maximum Nest
3. Error Components Logit
4. Nested logit Aggregate Lambda
5. Nested logit Lambda 0.35
6. Nested logit Lambda 0.05

Methods 1, 2 and 3 were estimated using Sawtooth Software CBC/HB. Methods 4, 5 and 6 are estimated using custom HB in R.

2.0 POST HOC MAXIMUM NEST

As the name of the method indicates this method is applied post hoc. The utilities are estimated without the nest information, which will be applied in the share calculation.

The first step we need to take is to determine which products belong together in nests. See section 6 for an example of how to do this. In the example below we have a simulator with 6 products and a None. Products 1, 2 and 3 belong to nest 1, products 4, 5 and 6 belong to nest 2 and the none is a nest on its own.

The second step is to compute the exponentiated utility for each product e^U . But now instead of proportioning all products like we do in a standard multinomial logistic, we compute the *maximum* exponentiated utility within each nest. This will always be a positive value since it is the exponentiated utility.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-4	1	0.8	0.6	1	1	0.71
<u>exp util</u>	0.02	2.72	2.23	1.82	2.72	2.72	2.03
Max <u>util</u> within nest		2.72			2.72		2.03

In the third step we will calculate what we will call **nest shares**. These nest shares are simply determined by the ratio of the maximum exponentiated utilities. So in our example the share of Nest 1 is $2.72/(2.72 + 2.72 + 2.03)$.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-4	1	0.8	0.6	1	1	0.71
<u>exp util</u>	0.02	2.72	2.23	1.82	2.72	2.72	2.03
Max <u>util</u> within nest		2.72			2.72		2.03
share for each nest		36.4%			36.4%		27.2%

Step four is to calculate the share of preference within each nest. This is done treating each nest separately. For each nest we compute the relative share using the standard MNL rule of e^U divided by the sum all those within the nest. These shares within the nest will sum to 100%.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-4	1	0.8	0.6	1	1	0.71
<u>exp util</u>	0.02	2.72	2.23	1.82	2.72	2.72	2.03
Max <u>util</u> within nest		2.72			2.72		2.03
share for each nest		36.4%			36.4%		27.2%
	0.37%	54.78%	44.85%	25.10%	37.45%	37.45%	
actual shares	0.13%	19.93%	16.32%	9.13%	13.63%	13.63%	27.23%

Finally, we multiply the shares within a nest by the nest share. This is shown on the last line above. The shares within each nest will sum to the nest share we computed. And the shares in the table below we compare the results from Post Hoc maximum nest share with regular share of preference.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-4	1	0.8	0.6	1	1	0.71
exp util	0.02	2.72	2.23	1.82	2.72	2.72	2.03
Max util within nest		2.72			2.72		2.03
share for each nest		36.4%			36.4%		27.2%
actual shares	0.37%	54.78%	44.85%	25.10%	37.45%	37.45%	
	0.13%	19.93%	16.32%	9.13%	13.63%	13.63%	27.23%
using "regular share of preference"	0.14%	19.07%	15.64%	12.76%	19.07%	19.07%	14.24%
shares for each class		34.85%			50.91%		14.24%

Nest 2 has three relatively strong products. Using regular share of preference the total for those three products is 50.91%. While using the maximum nest method it is only 36.4% because the share for that nest is based on its best item.

When we start simulating market changes more changes will appear. In the example below we increased the preference for product “A1” (by, for example, lowering its price or adding a feature). The product utility for product “A1” went up from -4 to -0.5.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-0.5 ↑	1	0.8	0.6	1	1	0.71
exp util	0.61 ↑	2.72	2.23	1.82	2.72	2.72	2.03
Max util within nest		2.72			2.72		2.03
share for each nest		36.4%			36.4%		27.2%

As product “A1” does not exceed the maximum within Nest 1 the nest shares are not changing, meaning we are not sourcing from Nest 2 or the none but just within Nest 1.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-0.5 ↑	1	0.8	0.6	1	1	0.71
exp util	0.61 ↑	2.72	2.23	1.82	2.72	2.72	2.03
Max util within nest		2.72			2.72		2.03
share for each nest		36.4%			36.4%		27.2%
actual shares	10.93% ↑	48.97% ↓	40.10% ↓	25.10%	37.45%	37.45%	
	3.98% ↑	17.82% ↓	14.59% ↓	9.13%	13.63%	13.63%	27.23%

If we look at the regular share of preference calculation based on IIA we see that product “A1” sources from all other products.

	Nest 1			Nest 2			None
	prod A1	prod A2	prod A3	prod B1	prod B2	prod B3	None
utility	-0.5 ↑	1	0.8	0.6	1	1	0.71
exp util	0.61 ↑	2.72	2.23	1.82	2.72	2.72	2.03
Max util within nest		2.72			2.72		2.03
share for each nest		36.4%			36.4%		27.2%
actual shares	10.93% ↑	48.97% ↓	40.10% ↓	25.10%	37.45%	37.45%	
	3.98% ↑	17.82% ↓	14.59% ↓	9.13%	13.63%	13.63%	27.23%
using "regular share of preference"	4.09% ↑	18.31% ↓	14.99% ↓	12.28% ↓	18.31% ↓	18.31% ↓	13.70% ↓
shares for each nest		37.39% ↑			48.90% ↓		13.70%

3.0 ERROR COMPONENTS LOGIT

The error components logit adds variables to induce patterns of correlation across alternatives. We do this by coding additional binary (0/1) variables for each nest. We need to have a reference category, so if we have 3 nests, we have 2 additional variables. Using the previous examples with SKUs 1, 2, 3 in Nest 1, SKUs 4, 5, 6 in Nest 2, and None as a reference Nest 3 would look like this:

Resp	Concept	SKU	Price	Nest 1	Nest 2
1	1	5	1	0	1
1	2	1	3	1	0
1	3	6	2	0	1
1	4	3	2	1	0
1	5	2	3	1	0
1	6	None	0	0	0

The additional nesting variables inform the upper level covariance matrix of an HB model, and therefore should help create more correlation among alternatives in a nest. The utility for an alternative above would be $U_{sku} + U_{price} + U_{nest}$.

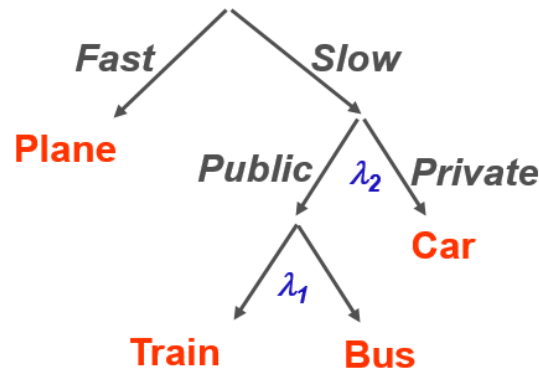
More details about Error Components Logit can be found in Train (2009), pages 139–141. One key point is that Train recommends the additional nesting parameters be *fixed* coefficients. This means they would be the same for all respondents, rather than respondent specific. However, our objective in this paper was to test another method that could be easily estimated with Sawtooth Software tools. Since Sawtooth Software does not allow us to specify fixed parameters, we specified the parameters as random.

We had hoped to find some improvement over standard HB by estimating all the parameters (include the nests) as random. Unfortunately, our results for ECL show that it did not change sourcing much at all from standard HB. We believe this is because we estimated the nesting parameters at the respondent level. So we call our results ECL*, with the intent of the * to indicate we did not follow Train’s recommendation to fix the nesting parameters.

We agree with Train that estimating the nesting parameters as fixed global parameters is most likely the best way to apply ECL. And we adopt this approach in the next section for nested logit, where we estimate global fixed parameters for the nests.

4.0 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 λ_1 represents the degree of similarity between train and bus.

We typically define the λ parameters in the interval $(0,1]$. When $\lambda = 1$, 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. One practical limitation is that as λ moves toward 0, numerical overflow can happen. This also depends upon the size of the utilities. For this reason we typically estimate $1/\lambda$ and constrain $1/\lambda$ in $[1,10]$ or $[1,5]$. The latter keeps $\lambda > .2$ which we have found to be aggressive enough for all practical purposes.

There are a few variations of nested logit, but we use 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.

Lattery (2016) describes the mathematical details of nested logit more fully.

5.0 ESTIMATING NESTED LOGIT WITH HIERARCHICAL BAYES

Nested logit extends the multinomial logit (MNL) model. To estimate nested logit, one must change the likelihood function from MNL to nested logit MNL. This is not something that can be done with Sawtooth Software, which hard codes the MNL likelihood function in the software. One approach to estimating nested logit is to use other estimation methods for nested logit like Empirical Bayes or Latent Class. Lattery (2015) shows that ensembles of Latent Classes can outperform HB estimation. Since nested logit is easy to implement with latent class, Lattery (2016) recommends applying nested logit by using ensembles of Latent Classes.

In this paper we wanted to directly compare post hoc buckets after standard HB estimation versus a nested logit estimation within HB. So we wanted to use HB estimation for both methods. We used custom code written in R that gives us full flexibility to modify likelihood functions and priors. It is analogous to what one would find with hierarchical Bayesian programs like WinBUGS, JAGS or Stan.

In addition to changing the likelihood function, we also need to specify how we should estimate the λ parameters. It is a complete disaster (in theory and practice) to simply include the λ parameters as additional respondent level utilities in the same upper level covariance with the other parameters. Lattery (2016) shows how including the λ parameters significantly lowered holdout log-likelihood from -5164.2 to -5828.2. In that case HB predicted holdouts significantly worse than a single Latent Class solution without nesting.

As mentioned above, we estimate $1/\lambda$ in the interval $[1,5]$. Our recommendation is that the $1/\lambda$ parameters be estimated at the global (not respondent) level. We also recommend that each $1/\lambda$ have its own normal prior that remains fixed during estimation. Our current default is for the base level of nests to use $1/\lambda \sim N(\mu = 3, \sigma = 1)$. The μ values allows relatively aggressive sourcing within the nest. At upper levels of the nest we may use μ values closer to 1, often choosing the aggregate level lambdas. Obviously μ and σ can be adjusted if desired, based on expert knowledge of expected sourcing. A better theoretical approach would treat μ and σ as random variables, and have another level of priors for μ and σ . But at the time of this writing this additional level of priors is still work in progress.

It is common for many marketing research firms to provide clients with Excel-based simulators for HB results. Excel is a convenient tool, but given the slowness of Excel we typically use point estimates for simulations, rather than draws from the posterior. Using point estimates in Excel is a practical limitation, especially when we have many parameters and large samples as we do in nested logit studies. In some cases using point estimates computed from the means of the posterior works very well. But with nested logit parameters, mean point estimates do a poor job of approximating the posterior. This is due to the significant impact that these parameters have on each draw, as they are exponents in the utility function.

The primary purpose of this paper is to compare the post hoc “bucket simulations” with comparable point-estimate predictions from full hierarchical Bayes models. Given our desire to use point estimates from full HB estimation, we decided to assume fixed λ values. We estimate the lambdas that come from an aggregate nested logit model and use those as our λ values for HB. And we also test a set of fixed lambda values: setting each λ at .35. Based on our experience a value of .35 tends to be aggressive, without being too aggressive and gives sourcing results comparable to the buckets approach described earlier in the paper. This makes the comparison of post hoc buckets vs. estimating nested logit more comparable. Using fixed λ parameters means we are using the same number of parameters as the post hoc buckets. It also removes any positive impact we have by modeling λ values.

6.0 DETERMINING WHICH ALTERNATIVES TO GROUP INTO BUCKETS OR NESTS

Any kind of nesting or buckets requires one to choose a specific structure for the alternatives. One can specify any structure they like. A marketing expert in a specific category may have a good idea what products tend to be more similar to others and define a structure based entirely on expert knowledge. Lattery (2016) describes one method for empirically deriving the nesting

structure. That method computes a “cross-purchase overlap” matrix for each pair of alternatives. It then treats those as pairwise distances and applies hierarchical clustering (Ward’s method). Those initial suggestions can be tested using aggregate nested logit to derive a final structure.

Since then we have also developed another method for empirically deriving nested structures. We first estimate a standard (no nesting) HB model to generate respondent level utilities (point estimates). For each alternative that we intend to put in a nesting structure, we compute the respondent-level utility for that alternative. Treating each respondent as a row and the utility of each alternative as a column, we compute the correlation matrix of the alternatives. We have found in several studies that this correlation matrix is very consistent with the structure we would have derived from the hierarchical clustering in Lattery (2016).

In this paper, we present two case studies. In both of these studies, we derived the nesting structure using hierarchical clustering as suggestions, then tested structures using aggregate-level nested logit. For comparison, we computed the correlation matrix of the standard HB utilities. The correlations between the respondent level utilities are shown below for Study 1. The color coding indicates the first level nests we derived, with the boxes indicating the second level.

Study 1

NEST	SKU	11	12	22	24	13	26	18	1	2	9	10	6	7	8	5	3	4	23	29	27	28	19	25	20	21	17	16	14	15
1	11	1.00	0.94	0.73	0.73	0.58	0.77	0.32	0.66	0.73	-0.70	-0.65	-0.40	-0.41	-0.53	-0.38	0.00	-0.15	0.43	-0.41	0.33	-0.30	-0.53	-0.12	-0.56	-0.64	-0.55	-0.38	0.04	-0.12
1	12	0.94	1.00	0.66	0.64	0.60	0.74	0.39	0.73	0.79	-0.66	-0.59	-0.27	-0.28	-0.40	-0.25	0.15	-0.01	0.33	-0.51	0.24	-0.38	-0.61	-0.28	-0.63	-0.70	-0.60	-0.43	0.01	-0.19
1	22	0.73	0.66	1.00	0.93	0.52	0.92	0.16	0.44	0.62	-0.84	-0.76	-0.73	-0.68	-0.81	-0.79	-0.25	-0.44	0.85	-0.12	0.71	0.03	-0.33	0.23	-0.39	-0.49	-0.59	-0.46	-0.10	-0.20
1	24	0.73	0.64	0.93	1.00	0.43	0.81	0.00	0.29	0.45	-0.77	-0.73	-0.83	-0.79	-0.86	-0.81	-0.44	-0.58	0.83	0.04	0.71	0.14	-0.33	0.40	-0.23	-0.34	-0.38	-0.33	-0.03	-0.02
2	13	0.58	0.60	0.52	0.43	1.00	0.51	0.10	0.36	0.52	-0.83	-0.83	-0.16	-0.32	-0.44	-0.15	0.28	0.17	0.49	-0.67	0.30	-0.34	-0.15	-0.20	-0.41	-0.48	-0.48	0.00	0.57	0.15
2	26	0.77	0.74	0.92	0.81	0.51	1.00	0.40	0.67	0.82	-0.79	-0.67	-0.50	-0.44	-0.59	-0.62	0.01	-0.19	0.65	-0.32	0.59	-0.13	-0.41	-0.07	-0.61	-0.69	-0.79	-0.63	-0.26	-0.43
2	18	0.32	0.39	0.16	0.00	0.10	0.40	1.00	0.80	0.71	-0.13	-0.02	0.35	0.35	0.25	0.08	0.53	0.45	-0.22	-0.59	-0.35	-0.73	-0.46	-0.60	-0.56	-0.53	-0.66	-0.63	-0.40	-0.68
2	1	0.66	0.73	0.44	0.29	0.36	0.67	0.80	1.00	0.95	-0.37	-0.24	0.19	0.21	0.05	0.01	0.57	0.43	0.00	-0.70	-0.08	-0.64	-0.63	-0.67	-0.85	-0.85	-0.84	-0.73	-0.36	-0.68
2	2	0.73	0.79	0.62	0.45	0.52	0.82	0.71	0.95	1.00	-0.58	-0.45	0.02	0.04	-0.15	-0.15	0.50	0.32	0.22	-0.70	0.15	-0.54	-0.56	-0.57	-0.86	-0.89	-0.92	-0.70	-0.29	-0.62
3	9	-0.70	-0.66	-0.84	-0.77	-0.83	-0.79	-0.13	-0.37	-0.58	1.00	0.96	0.55	0.63	0.75	0.57	0.07	0.23	-0.78	0.40	-0.60	0.13	0.16	-0.11	0.32	0.43	0.55	0.17	-0.29	-0.05
3	10	-0.65	-0.59	-0.76	-0.73	-0.83	-0.67	-0.02	-0.24	-0.45	0.96	1.00	0.57	0.70	0.78	0.53	0.16	0.28	-0.78	0.35	-0.55	0.10	0.06	-0.21	0.17	0.29	0.40	0.02	-0.44	-0.23
4	6	-0.40	-0.27	-0.73	-0.83	-0.16	-0.50	0.35	0.19	0.02	0.55	0.57	1.00	0.93	0.92	0.90	0.81	0.88	-0.85	-0.48	-0.76	-0.49	-0.03	-0.79	-0.23	-0.13	-0.05	0.02	-0.07	-0.28
4	7	-0.41	-0.28	-0.68	-0.79	-0.32	-0.44	0.35	0.21	0.04	0.63	0.70	0.93	1.00	0.94	0.81	0.76	0.81	-0.86	-0.34	-0.66	-0.36	-0.06	-0.77	-0.30	-0.19	-0.08	-0.11	-0.30	-0.43
4	8	-0.53	-0.40	-0.81	-0.86	-0.44	-0.59	0.25	0.05	-0.15	0.75	0.78	0.92	0.94	1.00	0.85	0.64	0.74	-0.91	-0.22	-0.75	-0.30	-0.02	-0.65	-0.10	0.02	0.10	0.00	-0.25	-0.30
5	5	-0.38	-0.25	-0.79	-0.81	-0.15	-0.62	0.08	0.01	-0.15	0.57	0.53	0.90	0.81	0.85	1.00	0.65	0.75	-0.83	-0.38	-0.73	-0.35	0.04	-0.65	-0.10	-0.01	0.17	0.25	0.10	0.02
5	3	0.00	0.15	-0.25	-0.44	0.28	0.01	0.53	0.57	0.50	0.07	0.16	0.81	0.76	0.64	0.65	1.00	0.94	-0.49	-0.79	-0.40	-0.59	-0.26	-0.94	-0.65	-0.58	-0.54	-0.29	-0.07	-0.51
5	4	-0.15	-0.01	-0.44	-0.58	0.17	-0.19	0.45	0.43	0.32	0.23	0.28	0.88	0.81	0.74	0.75	0.94	1.00	-0.60	-0.71	-0.54	-0.57	-0.19	-0.90	-0.49	-0.41	-0.36	-0.16	-0.01	-0.41
6	23	0.43	0.33	0.85	0.83	0.49	0.65	-0.22	0.00	0.22	-0.78	-0.78	-0.85	-0.86	-0.91	-0.83	-0.49	-0.60	1.00	0.14	0.81	0.31	-0.02	0.54	0.02	-0.10	-0.27	-0.15	0.16	0.13
6	29	-0.41	-0.51	-0.12	0.04	-0.67	-0.32	-0.59	-0.70	-0.70	0.40	0.35	-0.48	-0.34	-0.22	-0.38	-0.79	-0.71	0.14	1.00	0.29	0.78	0.27	0.77	0.67	0.68	0.63	0.19	-0.21	0.25
6	27	0.33	0.24	0.71	0.71	0.30	0.59	-0.35	-0.08	0.15	-0.60	-0.55	-0.76	-0.66	-0.75	-0.73	-0.40	-0.54	0.81	0.29	1.00	0.61	-0.01	0.45	-0.07	-0.15	-0.27	-0.19	-0.03	-0.02
6	28	-0.30	-0.38	0.03	0.14	-0.34	-0.13	-0.73	-0.64	-0.54	0.13	0.10	-0.49	-0.36	-0.30	-0.35	-0.59	-0.57	0.31	0.78	0.61	1.00	0.37	0.62	0.43	0.41	0.39	0.19	-0.11	0.23
7	19	-0.53	-0.61	-0.33	-0.33	-0.15	-0.41	-0.46	-0.63	-0.56	0.16	0.06	-0.03	-0.06	-0.02	0.04	-0.26	-0.19	-0.02	0.27	-0.01	0.37	1.00	0.34	0.45	0.45	0.45	0.64	0.28	0.46
7	25	-0.12	-0.28	0.23	0.40	-0.20	-0.07	-0.60	-0.67	-0.57	-0.11	-0.21	-0.79	-0.77	-0.65	-0.65	-0.94	-0.90	0.54	0.77	0.45	0.62	0.34	1.00	0.72	0.65	0.55	0.35	0.19	0.55
7	20	-0.56	-0.63	-0.39	-0.23	-0.41	-0.61	-0.56	-0.85	-0.86	0.32	0.17	-0.23	-0.30	-0.10	-0.10	-0.65	-0.49	0.02	0.67	-0.07	0.43	0.45	0.72	1.00	0.98	0.80	0.58	0.28	0.63
7	21	-0.64	-0.70	-0.49	-0.34	-0.48	-0.69	-0.53	-0.85	-0.89	0.43	0.29	-0.13	-0.19	0.02	-0.01	-0.58	-0.41	-0.10	0.68	-0.15	0.41	0.45	0.65	0.98	1.00	0.82	0.57	0.25	0.57
8	17	-0.55	-0.60	-0.59	-0.38	-0.48	-0.79	-0.66	-0.84	-0.92	0.55	0.40	-0.05	-0.08	0.10	0.17	-0.54	-0.36	-0.27	0.63	-0.27	0.39	0.45	0.55	0.80	0.82	1.00	0.75	0.35	0.72
8	16	-0.38	-0.43	-0.46	-0.33	0.00	-0.63	-0.63	-0.73	-0.70	0.17	0.02	0.02	-0.11	0.00	0.25	-0.29	-0.16	-0.15	0.19	-0.19	0.19	0.64	0.35	0.58	0.57	0.75	1.00	0.68	0.88
8	14	0.04	0.01	-0.10	-0.03	0.57	-0.26	-0.40	-0.36	-0.29	-0.29	-0.44	-0.07	-0.30	-0.25	0.10	-0.07	-0.01	0.16	-0.21	-0.03	-0.11	0.28	0.19	0.28	0.25	0.35	0.68	1.00	0.76
8	15	-0.12	-0.19	-0.20	-0.02	0.15	-0.43	-0.68	-0.68	-0.62	-0.05	-0.23	-0.28	-0.43	-0.30	0.02	-0.51	-0.41	0.13	0.25	-0.02	0.23	0.46	0.55	0.63	0.57	0.72	0.88	0.76	1.00

As shown above, the items within a nest are highly correlated. This is also true for Study 2. So in our case standard HB utilities without nesting are already picking up much of the correlation across alternatives. This is why respondent-level HB is so much better than aggregate models for sourcing. Correlation among utilities is also very important from a practical standpoint of applying the post hoc buckets. Post hoc buckets create a difference between what we are **simulating** (buckets) and what we are **estimating** (no buckets). The more closely post hoc simulation buckets align with the estimation correlation matrix, the more consistent simulations will be with estimations. If one is applying post hoc buckets it is important to check the correlation matrix of your standard HB utilities. If the correlation structure differs greatly from post hoc buckets, consider changing your buckets or dropping them.

7.0 CASE STUDY 1

Our first case study involves 29 SKUs in the paper towels category. Each screen was a virtual shelf showing 22 of the 29 SKUs at varying price levels. This was a volumetric study in that each respondent could choose as many products as they wanted on each screen (or none). For this paper, we recoded these volumes into an allocation. 1157 respondents completed the survey, with each respondent evaluating 16 shelves.

We empirically structured the 29 SKUs into 8 nests or buckets. We also had a second level of buckets that grouped nests {1,2}, {3,4,5}, {6}, and {7,8}. The nested logit and post hoc buckets both changed the sourcing in ways that we wanted. For example, we can start with a base case like this:

Brand	SKU	Price	Nest
Brand 1a	1 Large Roll, Full Size Sheets	\$1.09	1
Brand 1b	2 Bulk Rolls, Full Size Sheets	\$3.99	2
Brand 1a	1 Large Roll, Select A Size Sheets	\$3.29	2
Brand 1c	1 Giant Roll, Full Size Sheets	\$2.49	2
Brand 1a	12 Large Rolls, Select A Size Sheets	\$14.99	3
Brand 1a	6 Large Rolls, Full Size Sheets, Prints	\$11.99	4
Brand 1a	6 Large Rolls, Select A Size Sheets	\$11.99	4
Brand 1a	6 Giant Rolls, Full Size Sheets	\$10.49	4
Brand 1a	3 Mega Rolls, Select A Size Sheets	\$7.49	5
Brand 1a	4 Large Rolls, Full Size Sheets	\$6.49	5
Brand 2	6 Mega Rolls, Choose-A-Sheet	\$6.99	7
Brand 3	8 Large Rolls, Full Size Sheets, Prints	\$9.49	7
Brand 1b	4 Large Rolls, Full Size Sheets	\$4.49	8
Brand 1b	6 Giant Rolls, Select A Size Sheets	\$6.99	8
Brand 1b	8 Large Rolls, Full Size Sheets	\$8.99	8
None			

When we introduce a new item in Nest 7 for example, we should see more sourcing within Nest 7. The following table shows how the base case shares changed (multiplicatively) when doing this:

Nest	Standard HB	Post hoc Max Nest	ECL*	Nested Logit Aggregate λ	Nested Logit $\lambda = 0.35$	Nested Logit $\lambda = 0.05$
1	0.99	1.00	0.98	0.99	0.99	1.00
2	0.97	0.99	0.98	0.97	0.98	0.99
2	0.99	1.00	0.99	0.99	1.00	1.00
2	0.97	1.00	0.97	0.98	0.99	1.00
3	0.98	1.00	0.98	0.98	0.99	1.00
4	0.97	1.00	0.97	0.98	0.99	1.00
4	0.97	1.00	0.97	0.98	0.99	1.00
4	0.96	0.99	0.96	0.97	0.98	0.99
5	0.97	0.99	0.97	0.98	0.99	0.99
5	0.96	0.99	0.96	0.97	0.98	0.99
7	0.95	0.85	0.95	0.89	0.84	0.84
7	0.84	0.62	0.80	0.71	0.60	0.52
8	0.95	0.99	0.95	0.96	0.97	0.99
8	0.94	0.98	0.94	0.95	0.97	0.99
8	0.92	0.98	0.91	0.93	0.95	0.98
None	0.97	0.99	0.97	0.97	0.98	0.99

All the methods show relatively little sourcing outside Nest 7. Respondent-level utilities really do help, especially when the utility correlations are consistent with the buckets. Even with the standard HB most of the sourcing comes from the items in Nest 7, which become 95% and 84% of their initial share. The Post Hoc buckets draw much more from within Nest 7. The nested logit's sourcing will depend upon the λ values. Smaller λ values like .05 make the sourcing within Nest 7 stronger. As shown above, λ values of about .35 show similar sourcing to the Post Hoc nest. The aggregate λ values were higher than .35, which is why they show less aggressive sourcing than the post hoc or $\lambda = .35$.

One advantage of the nested logit is that one can vary the degree of sourcing by setting different λ values. This provides a way of tuning sourcing that is not possible with post hoc buckets. Even though we have shown a λ of .05 in the table above, we typically do not recommend $\lambda < .2$, as there is little to be gained with smaller values and they can cause numerical overflow in the computations.

We did not find much improvement using ECL* versus standard HB. But as mentioned earlier, that may be because we did not estimate the nesting parameters as global fixed parameters. Our objective in using ECL* was to modify the ECL approach to have a method that could be estimated via Sawtooth Software, with some additional coding. Our conclusion is that ECL* does not work. ECL may work, but that requires custom coding and we prefer the flexibility of nested logit.

To evaluate the fit of the models, we have 2 different types of holdout tasks. First, we have a random holdout task, where we removed one task for each respondent. The experimental design was a Sawtooth-type design with hundreds of versions. So we have 1157 (1 per respondent)

different holdout tasks at the respondent level. Second, we have a fixed holdout task. This task is almost the same across respondents, in that the SKUs were the same but the prices varied.

	Measure	Perfect	Sawtooth HB	Post-Hoc Max Nest	ECL*	HB in R (Base)	Nested Logit Agg λ	Nested Logit $\lambda = 0.35$	Nested Logit $\lambda = 0.05$
Random Holdout	LL	-600.1	-1778.4	-1823.3	-1769.7	-1794.8	-1797.6	-1813.1	-2211.7
	Pct Cert	100.0%	61.1%	59.6%	61.4%	60.5%	60.5%	59.9%	46.8%
Fixed Shelf	LL	-594.2	-1809.4	-1862.2	-1801.1	-1799.1	-1769.5	-1771.6	-2043.3
	Pct Cert	100.0%	53.5%	51.5%	53.8%	53.9%	55.0%	55.0%	44.6%
Combined Fit	LL	-1194.3	-3587.8	-3685.5	-3570.7	-3593.9	-3567.0	-3584.7	-4255.1
	Pct Cert	100.0%	57.6%	55.8%	57.9%	57.5%	57.9%	57.6%	45.7%

At the combined fit level, none of the nesting methods have a big improvement over their non-nesting counterparts. Only two of the methods show a decline: the Post Hoc and the very aggressive $\lambda = .05$. The $\lambda = .05$ is simply too aggressive, and even though we ran estimation through this model, the utilities are not able to estimate a consistent model. The best λ depends upon the data. One should not just pick any λ .

The post hoc shows only slight declines. This is true for both the random holdout and the fixed shelf. This is less decline than we expected given that we are estimating one model, and then simulating a different model. Bear in mind that the correlation matrix of the estimation model is very consistent with the nests. So applying the post hoc nests was a kind of consistent exaggeration of the base model.

For the fixed shelf we also compared aggregate shares across the 1157 respondents. Comparing observed shares versus simulated shares we see the following R^2 and mean absolute error (MAE):

Measure	Sawtooth HB	Post-Hoc Max Nest	ECL*	HB in R (Base)	Nested Logit Agg λ	Nested Logit $\lambda = 0.35$	Nested Logit $\lambda = 0.05$
R^2	0.974	0.973	0.974	0.976	0.986	0.991	0.997
MAE	2.0%	2.5%	2.0%	1.9%	1.5%	1.3%	0.8%

The table above shows that the nested logit predicted aggregate shares much better. But the post hoc nests made aggregate share predictions worse than standard HB. Again this is only one holdout task, and more tasks would give a better picture of how typical a loss in MAE is for post hoc buckets. But regardless of how many empirical tests we run, the reality is that the post hoc nests will always be theoretically questionable because they use one method for estimation and another method for simulation.

8.0 CASE STUDY 2

Our second case study involves 23 SKUs in the oral care category. Each screen was a virtual shelf showing 14–15 of the 23 SKUs at varying price levels. This was also a volumetric study in that each respondent could choose as many products as they wanted on each screen (or none). For this paper, we recoded these volumes into an allocation. 1026 respondents completed the survey, with each respondent evaluating 10 shelves.

In addition, this survey used a “reverse dual none.” Respondents were shown the virtual shelf and first asked whether they would buy any of the products. Then they were asked which they would buy (counterfactually if they said they would not). Respondents who said they would buy were coded as a single task, with the virtual shelf and none. Respondents who said they would not buy were coded as two tasks:

1. 1st task as above, with the None option selected
2. 2nd task removes the None option and choice is items selected.

We empirically structured the 23 SKUs into 8 nests. We also had a second level of buckets that grouped nests {1,2}, {3}, {4,5,6,7,8}. The nested logit and post hoc buckets both changed the sourcing in ways that we wanted. Like the previous study, the correlation matrix of the standard HB utilities was very consistent with the buckets:

	SKU	23	8	4	1	2	3	6	5	7	9	11	10	12	13	19	20	16	17	18	14	15	21	22
1	23	1.00	0.71	0.76	0.90	0.89	0.10	-0.42	-0.33	-0.41	-0.41	-0.77	-0.72	0.13	0.05	0.17	-0.08	-0.54	0.08	-0.29	-0.37	-0.28	-0.06	-0.29
1	8	0.71	1.00	0.94	0.80	0.79	-0.16	-0.60	-0.43	-0.46	-0.38	-0.50	-0.66	-0.21	-0.09	0.00	-0.18	-0.32	0.14	-0.17	-0.28	-0.08	0.30	-0.12
1	4	0.76	0.94	1.00	0.88	0.87	-0.09	-0.60	-0.46	-0.49	-0.47	-0.61	-0.69	-0.05	0.02	0.07	-0.19	-0.37	0.13	-0.24	-0.26	-0.13	0.20	-0.23
1	1	0.90	0.80	0.88	1.00	0.96	0.09	-0.48	-0.39	-0.46	-0.46	-0.74	-0.66	0.09	0.04	0.21	-0.09	-0.56	-0.04	-0.39	-0.37	-0.29	0.03	-0.34
1	2	0.89	0.79	0.87	0.96	1.00	0.05	-0.56	-0.46	-0.53	-0.49	-0.70	-0.60	0.08	0.04	0.22	-0.06	-0.54	0.01	-0.33	-0.34	-0.27	0.05	-0.30
2	3	0.10	-0.16	-0.09	0.09	0.05	1.00	0.69	0.77	0.68	0.55	0.09	0.29	-0.24	-0.43	-0.11	-0.13	-0.62	-0.71	-0.64	-0.72	-0.80	-0.53	-0.47
2	6	-0.42	-0.60	-0.60	-0.48	-0.56	0.69	1.00	0.94	0.94	0.74	0.46	0.52	-0.23	-0.39	-0.23	-0.05	-0.22	-0.63	-0.34	-0.38	-0.48	-0.45	-0.18
2	5	-0.33	-0.43	-0.46	-0.39	-0.46	0.77	0.94	1.00	0.98	0.82	0.45	0.47	-0.41	-0.52	-0.31	-0.15	-0.29	-0.64	-0.40	-0.53	-0.58	-0.42	-0.23
2	7	-0.41	-0.46	-0.49	-0.46	-0.53	0.68	0.94	0.98	1.00	0.86	0.52	0.49	-0.43	-0.54	-0.33	-0.15	-0.20	-0.60	-0.33	-0.47	-0.51	-0.37	-0.17
2	9	-0.41	-0.38	-0.47	-0.46	-0.49	0.55	0.74	0.82	0.86	1.00	0.59	0.59	-0.62	-0.68	-0.42	-0.25	-0.03	-0.45	-0.18	-0.44	-0.41	-0.32	-0.09
3	11	-0.77	-0.50	-0.61	-0.74	-0.70	0.09	0.46	0.45	0.52	0.59	1.00	0.78	-0.47	-0.40	-0.52	-0.16	0.38	-0.12	0.31	0.02	0.11	-0.15	0.16
3	10	-0.72	-0.66	-0.69	-0.66	-0.60	0.29	0.52	0.47	0.49	0.59	0.78	1.00	-0.24	-0.26	-0.28	-0.13	0.24	-0.37	0.04	0.04	-0.05	-0.38	-0.08
4	12	0.13	-0.21	-0.05	0.09	0.08	-0.24	-0.23	-0.41	-0.43	-0.62	-0.47	-0.24	1.00	0.89	0.62	0.18	0.02	0.14	-0.12	0.56	0.25	-0.16	-0.26
4	13	0.05	-0.09	0.02	0.04	0.04	-0.43	-0.39	-0.52	-0.54	-0.68	-0.40	-0.26	0.89	1.00	0.48	0.02	0.25	0.35	0.03	0.77	0.50	-0.09	-0.26
5	19	0.17	0.00	0.07	0.21	0.22	-0.11	-0.23	-0.31	-0.33	-0.42	-0.52	-0.28	0.62	0.48	1.00	0.55	-0.35	-0.23	-0.46	0.19	-0.14	0.29	-0.05
5	20	-0.08	-0.18	-0.19	-0.09	-0.06	-0.13	-0.05	-0.15	-0.15	-0.25	-0.16	-0.13	0.18	0.02	0.55	1.00	-0.30	-0.17	-0.01	-0.05	-0.20	0.54	0.56
6	16	-0.54	-0.32	-0.37	-0.56	-0.54	-0.62	-0.22	-0.29	-0.20	-0.03	0.38	0.24	0.02	0.25	-0.35	-0.30	1.00	0.68	0.75	0.73	0.86	0.04	0.28
6	17	0.08	0.14	0.13	-0.04	0.01	-0.71	-0.63	-0.64	-0.60	-0.45	-0.12	-0.37	0.14	0.35	-0.23	-0.17	0.68	1.00	0.78	0.56	0.73	0.22	0.34
6	18	-0.29	-0.17	-0.24	-0.39	-0.33	-0.64	-0.34	-0.40	-0.33	-0.18	0.31	0.04	-0.12	0.03	-0.46	-0.01	0.75	0.78	1.00	0.43	0.67	0.25	0.63
7	14	-0.37	-0.28	-0.26	-0.37	-0.34	-0.72	-0.38	-0.53	-0.47	-0.44	0.02	0.04	0.56	0.77	0.19	-0.05	0.73	0.56	0.43	1.00	0.85	0.06	0.06
7	15	-0.28	-0.08	-0.13	-0.29	-0.27	-0.80	-0.48	-0.58	-0.51	-0.41	0.11	-0.05	0.25	0.50	-0.14	-0.20	0.86	0.73	0.67	0.85	1.00	0.16	0.24
8	21	-0.06	0.30	0.20	0.03	0.05	-0.53	-0.45	-0.42	-0.37	-0.32	-0.15	-0.38	-0.16	-0.09	0.29	0.54	0.04	0.22	0.25	0.06	0.16	1.00	0.75
8	22	-0.29	-0.12	-0.23	-0.34	-0.30	-0.47	-0.18	-0.23	-0.17	-0.09	0.16	-0.08	-0.26	-0.26	-0.05	0.56	0.28	0.34	0.63	0.06	0.24	0.75	1.00

To evaluate the fit of the models, we used leave-one-out holdouts. We removed one holdout task per respondent, estimated the models and predicted the holdout. We repeated with a 2nd holdout, and then a 3rd. The experimental design was a Sawtooth-type design with hundreds of versions. So we have 1026 x 3 = 3078 unique holdout tasks at the respondent level.

	Measure	Perfect	Sawtooth HB	Post-Hoc Max Nest	ECL*	HB in R (Base)	Nested Logit $\lambda=0.35$
Round 1	LL	-113.8	-1503.3	-1523.6	-1504.9	-1624.0	-1414.6
	Pct Cert	100.0%	50.4%	➡ 49.6%	50.3%	46.1%	53.5%
Round 2	LL	-109.7	-1464.2	-1481.0	-1475.4	-1465.1	-1402.3
	Pct Cert	100.0%	51.7%	➡ 51.1%	51.3%	51.6%	53.9%
Round 3	LL	-107.9	-1311.7	-1317.3	-1294.8	-1366.9	-1264.2
	Pct Cert	100.0%	57.1%	➡ 56.9%	57.7%	55.1%	58.8%

In each of the 3 rounds of holdout tasks, the post hoc nest had slightly poorer fit than the standard HB model. The decrease in performance is small enough that one might choose the post hoc buckets anyway since they give better-looking sourcing. However, the big news in these results is that the nested logit with $\lambda = .35$ significantly improved fit in each of the 3 cases.

9.0 CONCLUSIONS

With respect to sourcing, respondent-level standard HB is a significant improvement over aggregate logit. While sourcing remains IIA at the respondent level, the correlations in utilities across respondents induce correlation among the alternatives. The method of post hoc nests is a simple approach to accentuate this nesting. But it means our simulations are a different model from our estimation.

The degree to which this theoretical problem between estimation and simulation converts to practical problems will depend upon your specific data and how consistent the correlation matrix of standard HB utilities is with your nesting structure. If one is applying post hoc buckets it is important to check the correlation matrix of your standard HB utilities. If the correlation structure differs greatly from post hoc buckets, consider changing your buckets or dropping them. In our two case studies the post hoc nests were very consistent with the utility correlations. Applying post hoc buckets only slightly lowered the holdout fit.

Applying nested logit to the estimation process requires much more technical expertise. It cannot be done with standard software including Sawtooth Software. We used a custom program in R. This allowed us to change the likelihood function. Nested logit has the additional advantage that one can empirically tune each of the λ parameters, while post hoc buckets have no such parameter. To make fairer comparisons, we use simple λ parameters, setting each nest to have λ of .35. This performed very well in our two case studies. It improved the respondent-level fit significantly in study 2, and the aggregate fit in study 1. Of course, there is room for improvement by modeling the λ parameters, rather than setting them at .35. We discussed how this can be done in the nested logit section.

We do not recommend post hoc nests for every study. First, one should run a standard HB estimation, and check the sourcing. If the standard HB sourcing is not enough, then compute the correlation matrix of the respondent utilities. If the post hoc buckets are defined consistently with those correlations, it is likely the resulting simulations are also relatively consistent with the

model estimated. And the simplicity of applying post hoc buckets vs. the complexity of other methods makes it an alternative worth considering when standard HB sourcing is not enough.



Kevin Lattery



Jeroen Hardon

REFERENCES

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

Lattery, Kevin: The Art and Science of Nested Logit: Case Studies from Modeling Many SKUs. 2016. *Proceedings of the Sawtooth Software Conference*, pp 281–294.

Train, Kenneth: Discrete Choice Methods with Simulation. 2009. *Cambridge University Press*.

PREDICTIVE ANALYTICS WITH REVEALED PREFERENCE/ STATED PREFERENCE MODELS

PETER KURZ

KANTAR TNS

STEFAN BINNER

BMS MARKETING RESEARCH + STRATEGY

MOTIVATION FOR THIS PAPER

In fast moving consumer goods, market simulations based on Conjoint Analysis/DCM have some model limitations which are well known among researchers and end users. The main limitation is the lack of real market circumstances such as awareness, distribution and out of stock effects because choice models are not able to capture such effects due to the synthetic interview situation.

In order to adjust choice models to these market circumstances and therefore better interpret the results, researchers often apply revealed preference data (past data from sources such as sales data, scanner or household panels) to the choice data. Usually this revealed preference data represents the average of a certain period of time (e.g., last year, until now) in a certain distribution channel. However, if we look into the detailed data points within any given time period, we can see that volume sold, or market shares of single SKUs can vary dramatically over time:

Figure 1. Time series for single product over the period of 157 weeks.



Looking at the above times series represented in Figure 1, one could wonder if conjoint results are really free of context effects. Is the derived share of choice really not related to the point in time a study was conducted, i.e., at moments of intense advertising, social media reaction or holiday season?

Hypothesis: The point in time when the survey was conducted might have a significant influence on the model results, i.e., shares of choice or price elasticities.

The above hypothesis does not address “simple context effects” like different weather conditions that might influence the choice behavior. Many papers show that conjoint models are relatively stable against such effects. It is rather the general market situation in a certain time period in which the conjoint interviews were conducted.

In order to understand the possible influence on model results, it seems quite desirable to apply revealed preference data (past data from sources such as sales data, scanner or household panels) in order to improve survey-based models. For “Price Only Discrete Choice” models it might be furthermore useful to include observed market reactions (e.g., past effects of price

increase or promotions) from revealed preference data, in order to better simulate and predict pricing scenarios in the future.

The erratic increase of data (within the last 10 years the amount of data increased by factor of 1,000 and latest developments such as IOT [internet of things] or smart cities will ensure that this trend will continue), larger and scalable computation power and improved forecast models provide new opportunities to combine survey-based models with revealed preference data.

RPSP MODELS (REVEALED PREFERENCE/STATED PREFERENCE MODELS)

Depending on the objectives of predictive analytics, RPSP models may consist of up to three data sources:



The interview data from the conjoint/DCM exercise are defined as the stated preference data (SP) and represent the basis for simulating “what-if scenarios” in the marketplace. The revealed preference data (RP) are the past market actions recorded in scanner panels, sales and other data collected in the market reality. The stated preference data allow to simulate situations which never have been seen in the real world so far e.g., introduction of new products or pricing strategies which haven’t been tested in the real market. Revealed preference data are perfect representations of the market’s past and report all actions which could be observed in the real world. On that basis, past price changes or introduction of products could be modeled. The third source which could be included in analysis are social media data. With this data we get insights in the reaction of consumers on past actions. For example, how price changes were recognized and perceived or how customers communicate about promotional campaigns or discounts.

INDICATION FOR RPSP MODELS

SP Models are widely used in market research and have a great academic background. We all believe (and a large number of validation studies showed) that Conjoint/DCM provides reliable information about preferences and elasticities in the simulation model. Price-only discrete choice models mimic the marketplace pretty well and are close to decisions respondents make every day when buying products. No other methodology allows to simulate and compare the effect of several new products which are currently not available in the marketplace and allows to simulate the effect of price changes that were never seen in the real market. But due to the lack of awareness and distribution information the results could be different from market data. RP data are closer to market reality, but do not allow simulating the future. Therefore, the desire of combining the two information sources into one large model is quite logical.

However, in order to combine the SP model with RP data, we need a basic understanding about similarities and differences between revealed and stated preference!

Why do we expect that stated and revealed preferences show different results?

Beside volume, distribution, and out of stock effects there further might be a difference in the way consumers make their choices:

- Sales and Store Data (RP):
 - High complexity of “choice task” (real decision) and many alternatives (e.g., large and sometimes incomplete shelves, shop assistants at the counter, second placements at check out)
 - Varying context (e.g., different retail channels or store types)
 - Time pressure, distractions (e.g., shopping trip during lunch break, stop at a drive-in)
 - Budget constraints (e.g., end of the month)
 - Cross-category-decision (e.g., What should I cook today?)
- Price-only Discrete Choice (SP):
 - Clear description of attributes and levels (creating 100% awareness)
 - Lower motivation to answer the exercise carefully (e.g., panel burnout)
 - ICT (Individual Choice Task Threshold: lower attention after too many choice tasks)

Before planning an RPSP model in a specific marketplace one should therefore try to answer the following questions:

1. *Would we expect that stated and revealed preferences show different results?*
If we assume that both data sources deliver the same results, we better stay with RP-data.
2. *Could we identify if stated and revealed preferences show different effects?*
Before combining RP-data we should look into our SP-data and identify the effects we can model with this data. Usually, many of the effects we see in our SP-data could also be seen in the RP-data. Only if we identify some effects in the RP-data that could not be modeled in our SP-Model we should think about combining them. One should always keep in mind, that a combined model has a much larger complexity and therefore needs to be a significant added value from including the additional RP effects.
3. *Is the available RP data source sufficient to separate and explore preferences from context effects and constraints (e.g., availability, out of stock)?*
Most RP-data show effects which could not be modelled by SP-data. Unfortunately, these effects can also not be isolated in the RP data. A peak in the RP-data may be caused by a combination of circumstances which occurred at the same point of time. Such overlay of, e.g., price changes, advertising campaigns, press coverage or others could make the RP-data quite useless for forecasting purposes. RP-data which can really be decomposed to see single effects are rather rare.

If the answers to these questions are not positive, one should proceed with the SP-Model only, thus avoiding the complexity of adding the two data sources.

If the underlying decision process is different for SP and RP data and it matters for our predictions, then one should consider combining the two data sources.

SOURCE OF RP-DATA

Time series data (sales, scanner data, panels etc.) provide different levels of depth and insights on a single product (SKU) level:

- Retail scanner data or sales data might only show volumes (units or value) of the clients' product at different moments of time and at different prices.
- Time series from a company with its own distribution channel (e.g., fast-food chain) can record market shares of all offered SKUs including information about the price and other factors such as outlet type or region.

For those products with price changes during the recorded period, price- and cross-price-elasticities based on the RP-data can be derived.

COMPONENTS OF RP-DATA

A useful abstraction for selecting forecasting methods is to separate a time series into systematic and non-systematic components. Systematic components of the time series show consistency and/or recurrence and can therefore be described and modeled. Non-Systematic components of the time series are the ones that cannot be directly modeled or explained.

A given time series usually is thought to consist of three systematic components: level, trend and seasonality. In addition, there is one non-systematic component called "noise."

Level is the average value in the series. That could be best described as the baseline without any time-dependent change over the period of the data.

Trend describes increasing or decreasing values over time.

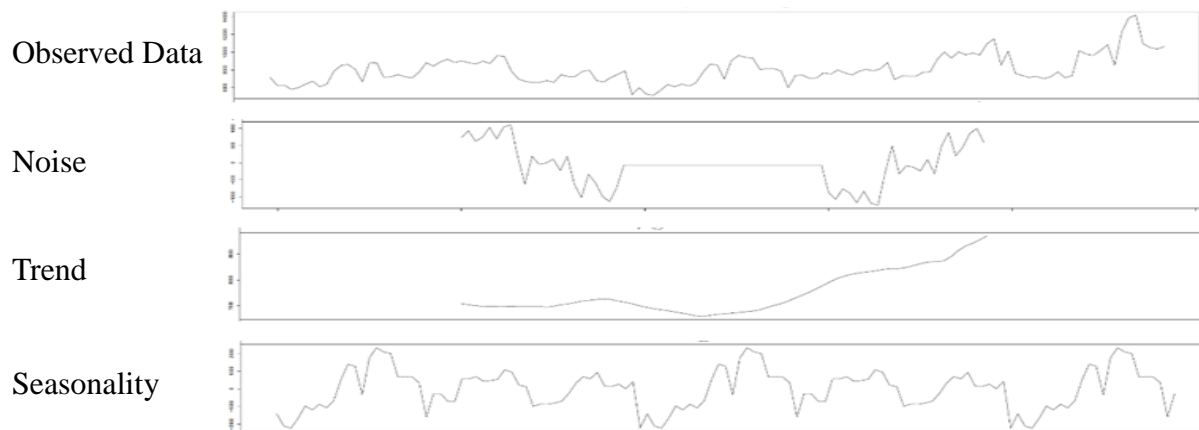
Seasonality describes cyclic effects that can be iterative and observed during the measurement period. For example different buying behavior during the summer season which occurs every year. This component could be isolated from the baseline (Level) and an overall increase or decrease over the complete period (Trend) and represents short-term cycles within the time-series.

Noise is the non-systematic component in each time series, also called the random variation. This component is the share which could not be described by the actual model used to decompose the components.

SEPARATION OF RP-DATA COMPONENTS

In time series analysis we usually try to identify the systematic components by separating the level, general trend, seasonality and the noise from the observed data on SKU level.

Figure 2. Decomposition of a time series for single SKU over past 157 weeks.

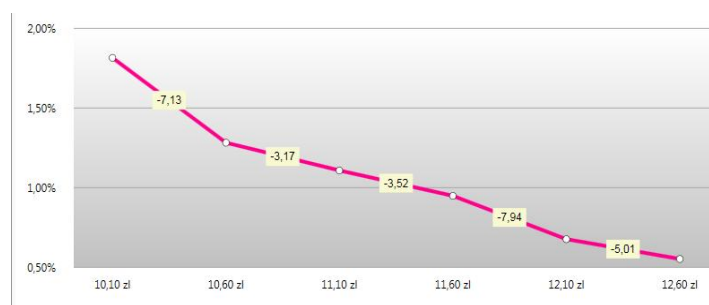
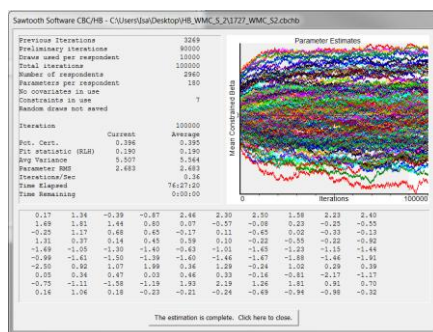


One of the most common models for seasonal decomposition of time series is the LOESS model described by Robert Cleveland, William Cleveland, Jean McRae and Irma Terpenning in the early 90s. For our RP-data in this paper we applied this model for decomposing level, trend, seasonality and noise and to apply the components to our SP-data.

NOISE (RANDOM COMPONENT) IN RP-DATA

The random component is a very important part of the decomposition. It shows the amount of unexplained information in the data. The lower the random component in the model, the better the decomposition. The random component can therefore be used to compare and benchmark different decomposition models in order to find the best fit (lowering the random component). The goal of each decomposition is to retrieve a small random component so that the systematic components explain most of the data. Furthermore, one can calculate the confidence intervals from the size of the random component in order to understand how well trend and seasonality fit.

STATED PREFERENCE DATA



Stated preference data is usually derived through a conjoint or discrete choice model and the part-worth utilities estimated from this experiment. Based on either point estimates, draws or an upper-level model, market scenarios can be modeled, and consumers' reactions simulated (e.g., to changes in pricing of individual SKUs). Most models in everyday work are using pseudo-individual estimates based on hierarchical Bayes regression.

The results of the what-if scenarios simulated based on part-worth models are the basis for combining SP- and RP-Data. Therefore, it is crucial to derive valid SP-models as well as reliable and meaningful what-if scenarios. Only if the simulated scenarios are comparable to the real market situation one can take out advantage of combined models.

LEVEL OF AGGREGATION IN SP-DATA

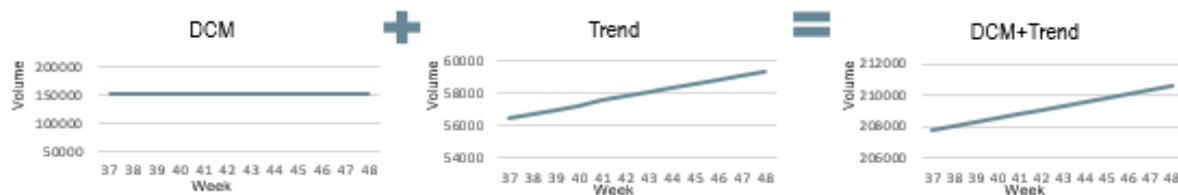
Choice experiments such as Conjoint/DCM allow to derive cross-price elasticities and preference information for a large number of possible scenarios on a nearly individual level. In contrast, most of our RP-sources have only aggregate-level data. And usually even if we have individual information in our RP-data we seldom have the same individuals in our SP-data. Therefore, in order to compare the SP-data with aggregated RP-data we need some kind of aggregation of the SP-data (for instance Channel, Store, Customer Segment). However, we know from previous studies that “the ignorance of heterogeneity in the aggregate model leads to biased forecasts” (Feuerstein, Natter, Kehl 1999). When merging the RP- and SP-models together one should therefore take great care about a proper aggregation level. The closer we can bring our two data sources and the closer we are to the individual level, the better the two models could fit and correct the SP-data in a meaningful way. The results of the study from Feuerstein, Kehl and Natter (1999) showed, “that scanning data are useful to improve external validity of CBC models by introducing dynamics of the static conjoint models . . .” The real advantage of RPSP-models is the dynamic from the RP data that enhance the static SP data with this component.

COMBINING TIME SERIES

General Trend

From our choice model we can derive the shares of choice for an actual market scenario. For each SKU within this scenario we can then use the respective decomposed RP-sales data (level, trend, seasonality). The level is used to calibrate the share of choice from the base case (actual market simulation) to sales units. The decomposed general trend from the SKUs’ sales data is used to modify the units for the forecasting period (e.g., 12 weeks). This results in (12) different scenarios which show the influence of the trend in the forecasting period.

Figure 3. Volume + Trend = Volume Sales
corrected for the 12 weeks forecasting period 37/17–48/17;
Cheeseburger @ Price 1.10€/Share of Choice 12.35%



Result of this correction is a forecast for the development of the unit sales for the 12 weeks forecasting period.

Seasonal Component

Applying the seasonal component of the sales data allows further correction of the forecasting period with the seasonality effect. The approach is very simple, because one only needs to apply the extracted time series component on top of the trend corrected unit sales.

Figure 4. Volume + Trend + Seasonality = Volume Sales corrected for the 12 weeks forecasting period 37/17–48/17; Cheeseburger @ Price 1.10€/Share of Choice 12.35%

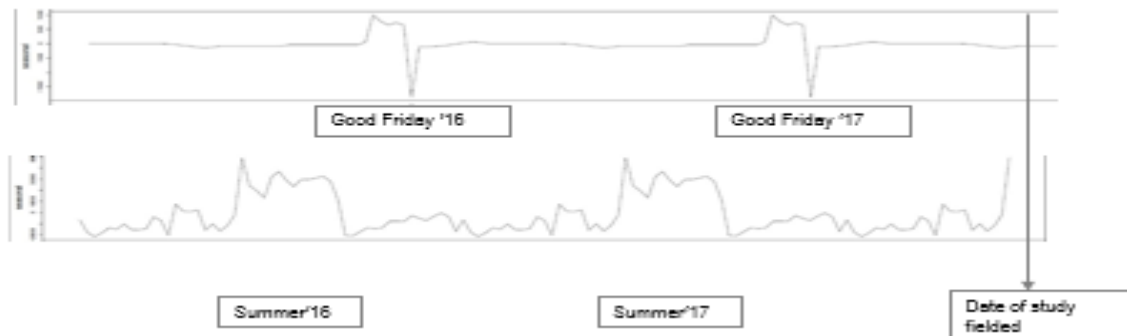


After this correction the result is a dynamic simulation model for the 12-week forecasting period. We don't assume any longer that the static results from the choice model are sufficient in order to forecast the market because the integrated dynamics from the RP-data results in a much more realistic time course.

Special Effects

The challenge is not to use all information. Some seasonal effects are decomposed on periodic events in the past, which have only an effect for exactly the same event in the future!

Figure 5. Seasonal-effect for sales of Sandwiches in a Drive-In Restaurant at a highway over the 157 weeks' time period.



This doesn't cause problems if the event is repeated in every period and its occurrence can be predicted, like our example of Good Friday, where consumers buy fewer sandwiches with meat on this day of the year. But if we see periodical effects like traffic jams that cause higher sales at fast food restaurants close to highways, we never would know, if such an event occurs during our forecasting period and if, when.

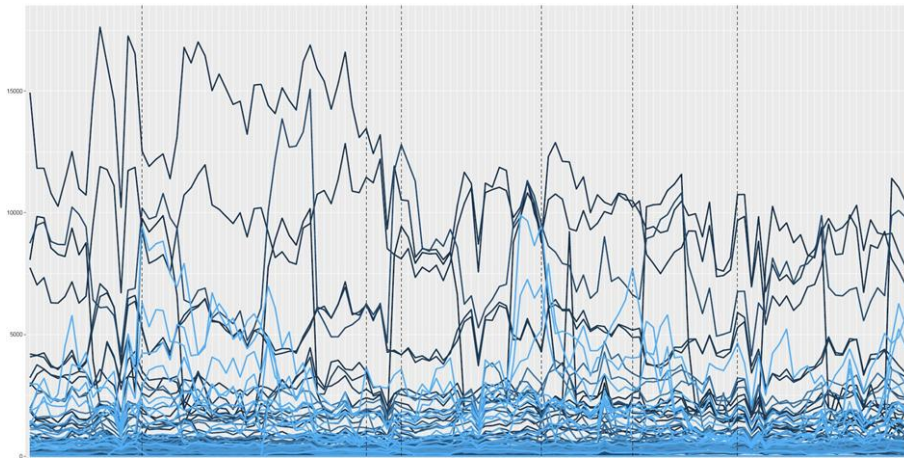
HANDLING A HUGE AMOUNT OF INFORMATION

Typically, one ends up with a very large number of time-series, at least with one series for each product/SKU. However, in addition there are often also different sales channels, store types,

geographical areas and locations. As a result, there are multiple different trend and seasonality components.

Imagine the fast food industry: A drive-in close to a highway has a completely different profile (holiday season, traffic jams, . . .) than a restaurant at a downtown area or an outlet in a shopping mall. The reduction of this complexity to an amount of information which could be handled is the challenge researchers face when integrating time-series models by deciding which systematic components should be included or not. Some periodic effects might be systematic or only random due to the usually short reference period used in market research. Usually one suggests having 40 to 50 cycles for analyzing seasonal effect. In market research we are happy if we have two to three years (e.g., 2 to 3 replicates) to run our models.

Figure 6. Sales figures for 97 different items over the 157 weeks time period.



Forecast on Sales: In Statistic Markets RP Data Fits Well

In most of our simulations we have no information about cannibalization, cross-selling or even how many customers we would lose completely when changing the current market environment. Time series could help to analyze past events and project them to the future, e.g., if there was a price change two years ago, one may predict from past customer reactions what will happen in the future.

**Figure 8. Sales figures for single SKU over the 157 weeks time period.
Identifying the two past price increases.**



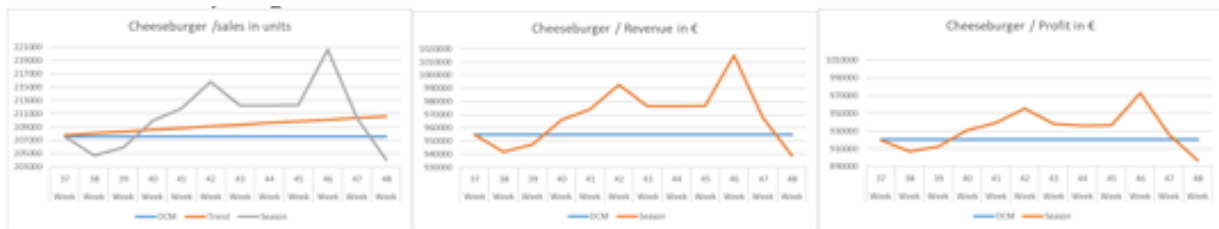
But there is no chance to forecast effects of future price changes if we did not observe comparable actions in the past. Also, if we introduce new products, we have no valid mean to

derive forecasts from the RP-data. Here we need the results from SP-models, which are based on survey data and therefore allow to simulate scenarios with new products, multiple price changes or even cannibalization between products.

FORECAST ON SALES/REVENUE/PROFIT

In the following example from a company with its own distribution channel, sales data including market shares of all offered SKUs, information about the price and other factors such as outlet type or region is available. This allows a forecast based on DCM simulation taking the decomposed RP data (market, trends and seasonal effects) for each single SKU into account in order to predict revenue and profit:

Figure 9. Sales in units based on DCM simulation (blue) and corrected for trend (orange) and seasonal effects (grey) (12 weeks period forecast) and multiplied by price 1.10€ = Revenue and subtract the cost of 0.52€ = Profit.



The static DCM shows the constant amount of sales throughout the whole period. Correcting with RP data we include dynamic into the forecast, evaluating the simulated performance under the assumption that trend and seasonality will be the same as in the past.

RPSP SIMULATIONS

Simulations of different market scenarios demonstrate the power of the RPSP approach.

Figure 10. SP and RPSP Sales and Profit under different pricing for Chickenburger and Cheeseburger (12 weeks period forecast).



Figure 10 shows the influence of two price changes for cheeseburger and two price changes for chickenburger during our simulated time period of 12 weeks. As price changes did not occur at the same time in the past, the cross elasticities for a simultaneous price change could not be estimated from the RP-data. This is the real strength of the RPSP-model.

GENERAL EVALUATION OF RPSP MODELS

- SP models have their strength in simulating dynamic markets.
- RP models are usually preferred in static markets.
- The strength of RPSP models is to combine accurate information from the past with the power of dynamic simulations.

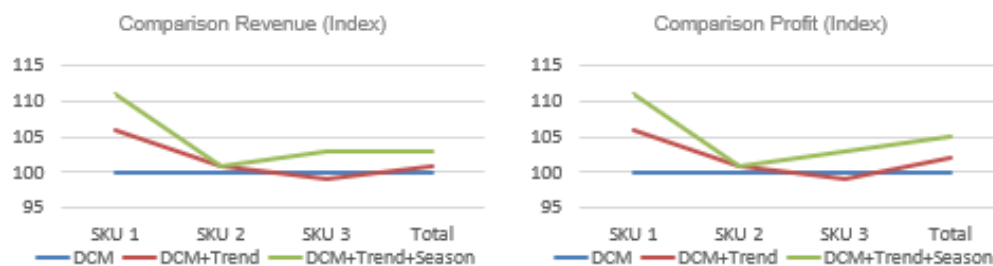
But,

- RPSP models are not suitable for “new to the world products,” due to the obvious lack of RP-data for these SKUs. In addition, there is often only limited data available for recently introduced products, so seasonal effects cannot be measured.
- RPSP-models usually do not gain much insights if we miss data on SKU and channel level. Aggregate sales data seldom improve the model when we only have them on an aggregated market level.
- Price only discrete choice models based on HB Draws can already drive super-computers to their edge of capacity. The combination of DCMs and RPSP models further multiply complexity if we include our RP-information in the upper level of the hierarchical Bayes estimation. The data challenge is resulting in an unsolvable optimization problem should we try to simulate multiple changes in multiple SKUs during our forecasting period. RPSP models therefore still need project specific adaptations in order to make them feasible.

Summary: Benefits of RPSP Models

The application of Time Series Corrections (RP) on our Share of Choice Simulations (SP) has a big impact, especially if we derive revenue or profit predictions:

Figure 12. Comparison for the different simulations (12 weeks period forecast).



Compared to the simulation based on the DCM only (share of choice)—the corrections for trend and seasonality (RP) lead to an improved prediction of revenue and profit. This shows the relevance of such corrections for business decisions. The application of RPSP models showed also a significant impact of the point in time on dynamic market simulations such as simulation of price reactions:

Figure 13. Comparison of Relative Change in Units and EoP over 12 weeks after correction for trend and seasonality.



If there are no data or resources for RPSP models, one should nevertheless consider the possible impact of the point in time the study was conducted. As we saw in this paper the point in time can have a significant impact on the predictions if we ignore time series information.



Peter Kurz



Stefan Binner

REFERENCES

- Brockwell, P.J.; Davis, R.A. (2016): Introduction to Time Series and Forecasting; Berlin, Heidelberg (Springer)
- Cleveland, R.; Cleveland, W.; McRae, J.; Terpenning, I. (1990): Seasonal Decomposition of Time Series by LOESS; The Journal of Official Statistics.
- Feuerstein, M.; Natter, M.; Kehl, L. (1999): Forecasting Scanner Data by Choice-Based Conjoint Models; Proceedings of the 1999 Sawtooth Conference
- Hardt, N.; Kim, Y.; Joo, M.; Kim, J.; Allenby, G.M. (2017): Reconciling Stated and Revealed Preferences; AMA ART Forum 2017, Presentation
- Hamilton, J.D. (1994): Time Series Analysis; Princeton (University Press)
- Huber, J. (1992): Comment on McLauchlan; Proceedings of the 1992 Sawtooth 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.
- Lütkepohl, H. (2006): New Introduction to Multiple Time Series Analysis; Berlin Heidelberg (Springer)

Miller, J.D.; Forte R.M. (2017): Mastering Predictive Analytics with R; Packt Publishing—
ebooks

Miller, T.W. (2014): Modeling Techniques in Predictive Analytics with Python and R: A Guide to
Data Science (FT Press Analytics)

THE PERILS OF IGNORING UNCERTAINTY IN MARKET SIMULATIONS AND PRODUCT LINE OPTIMIZATION

SCOTT FERGUSON

NORTH CAROLINA STATE UNIVERSITY

ABSTRACT

Quantitative market research models facilitate the creation of market simulators and the formulation of product line optimization solutions. Results from market simulators provide insight into how a population might respond to new product offerings, guiding decisions about product configuration and price. When physical product lines are created, the results from these simulations can also inform production and resource allocation decisions. The work presented in this paper highlights consequences of ignoring uncertainty associated with market-driven product line optimization problems, with a specific focus on parameter uncertainty. A two-objective optimization problem is introduced that maximizes revenue from the product line under a nominal model while also maximizing the worst case revenue from an uncertainty set of models. Here, the nominal model represents the mean of the posterior distribution of a hierarchical Bayes mixed logit model while the uncertainty set is represented by 800 draws from the posterior distribution. A third objective is also introduced that minimizes the variation of First Choice Share within the product line. The importance of this objective is demonstrated by illustrating the variation in share captured by each product when considering the models in the uncertainty set. This variation is discussed in the context of production and resource allocation decisions.

INTRODUCTION

Consider a manufacturer who is interested in creating a line of products for a heterogeneous market. The decision (design) variables for such a problem are product content (configuration) and product price. Configuration and pricing decisions can be informed by a market simulator that becomes the engine driving the product line optimization problem. Strategies for formulating and solving product line optimization problems have been presented at previous Sawtooth Software conferences [1–4], and even more references can be found in the literature [5–8]. These works have also shown that product line optimization problems are challenging for even modern optimization algorithms because they have large design spaces (billions or more possible combinations) and gradient-based optimization techniques cannot be used because of mixed-integer problem formulations.

The business objective for product line optimization problems is often revenue maximization, but the value of using objectives related to share of preference, profit, and commonality has also been demonstrated [4]. Once a solution has been found, decisions are made about product configuration, price, and production quantities. These outcomes are significant; manufacturers must order parts, design and construct assembly lines, and negotiate for shelf space. As noted by Bertsimas and Misis, product production decisions are both infrequent and require a commitment of manufacturer resources in a way that “cannot be easily reversed or corrected” [9].

There are many sources of uncertainty that, if not considered when solving the optimization problem, can translate to product line solutions with disastrous market performance. As

discussed in [9], at least two forms of uncertainty can be associated with the choice model: structural and parameter. Structural uncertainty can be thought of as demand model misspecification [10–12]. Parameter uncertainty is related to the model parameter estimates—including, but not limited to, part-worth values and segment probabilities. Additionally, uncertainty exists when considering competitor product configurations and prices, and the manufacturer’s own product attributes and component costs. In this paper, the focus is on the uncertainty in parameter estimates.

Optimization studies utilizing a single set of part-worth coefficient point estimates per respondent (such as the mean of the lower-level posterior distribution in a hierarchical Bayes mixed logit model) benefit from reduced computational cost. However, they neglect how the reported objective function is impacted by parameter uncertainty. Recognizing the potential hazards of using a single set of point estimates when simulating market behavior, especially if used to inform resource allocation decisions, researchers have proposed simulation strategies using draws from the posterior distribution, randomized first choice [13], interval variables, and moment estimation.

Building on these efforts, a robust revenue optimization approach has been introduced by Bertsimas and Misic that maximizes the worst case revenue of the product line under uncertainty. The work in this paper expands on their approach by reformulating the optimization problem as one with multiple objectives. The first objective maximizes overall revenue given a “nominal” model, while the second objective maximizes worst case revenue from an uncertainty set (of models). Realizing that the solution will also drive product inventory and manufacturing decisions, this paper introduces a third objective that considers the variation in choice amongst the products *within the product line*.

The approach presented in this paper is important because it highlights the value forfeited when uncertainty is ignored in product line optimization problems. By reformulating the optimization problem with multiple objectives, a decision-maker can develop a richer understanding of the tradeoffs (and risk) associated with different product line solutions. This work also demonstrates the inherent value of quantitative market research models and market simulators throughout the many stages of the design process.

DESCRIPTION OF RELEVANT LITERATURE

The papers listed in Table 1 provide a representation of how uncertainty has been addressed in recent product design literature. As stated in the previous section, these methods use draws from a posterior distribution, interval variables, or moment estimation.

Camm et al. [14] and Wang et al. [7] use samples from the posterior distribution and introduce post-optimality robustness tests that assess the negative impact of part-worth uncertainty. In [14], individual draws are used so that the deterministic optimization problem can be repeatedly solved. The optimal product configuration was also found using part-worth coefficient point estimates. Resultant solutions were then compared, and the product configuration that maximized first choice share (FCS) when using point estimates aligned with only 23.5% of the random draw solutions. Wang et al. [7] implemented a sample average approximation method using stochastic discrete optimization [15]. Parameter uncertainty was modeled by pulling multiple draws from a respondent’s posterior distribution. Each draw was then treated as a separate respondent, and the product line was optimized. Results from this study

showed that as the sampling of the posterior distribution increased, the number of optimal products reduced.

Wang and Curry [16], Luo et al. [17], and Besharati et al. [18] defined part-worths using interval variables and investigated the best and worst cases of product utility. Wang and Curry [16] studied robustness in the share-of-choice problem by assuming that individual preferences were bounded, independent, and symmetric. Also, the covariance matrix for individual level part-worths was assumed to have a diagonal form, preventing correlation among product features. Luo et al. [17] and Besharati et al. [18] used segment-level part-worth confidence intervals and calculated the lower and upper bounds of product utility. Both studies only considered the design of a single product (rather than a line) but considered multiple design objectives; namely, maximizing the share of preference using the nominal model, minimizing variation in share of preference, and minimizing the worst case performance. Resende et al. [19] advanced these studies by considering a profit objective and estimated the first and second moments of the objective function by applying the delta method [20]. A closed-form solution was then introduced using a Taylor series expansion when considering a multinomial logit model at a pre-specified risk level.

Table 1. Recent literature considering parameter uncertainty when using market research models in product (line) optimization.

Reference	Method to treat uncertainty in discrete choice methods	Design problem	Design variables	Design objective
Camm et al. [14]	Samples from posterior distribution	A single product	Discrete product attributes	Maximize FCS
Wang and Curry [16]	Manual definition of part-worth intervals	A single product	Discrete product attributes	Maximize FCS
Luo et al. [17]	Interval estimates of part-worths using 95% confidence levels	A single product	Discrete product attributes	Maximize nominal SOP, Minimize SOP variance, Minimize worst-case performance
Besharati et al. [18]	Interval estimates of part-worths using 95% confidence levels	A single product	Discrete product attributes	Maximize nominal SOP, Minimize SOP variance, Maximize engineering design performance
Resende et al. [19]	Moment estimation of market share based on continuous probability function of part-worths	A single product	Continuous product attributes	Maximize profit at specified downside risk tolerance
Wang et al. [7]	Samples from posterior distribution	Product line	Discrete product attributes	Maximize FCS
Bertsimas and Misic [9]	Samples from posterior distribution	Product line	Discrete product attributes	Maximize worst-case expected revenue

FCS: First Choice Share

SOP: Share of Preference

The recent publication by Bertsimas and Misis [9] most directly motivates the work in this paper. Product line robustness is explored by formulating an optimization problem that maximizes worst-case expected revenue over an uncertainty set, as shown in Equation 1.

$$\max_{S \subseteq \{1, \dots, N\}: |S|=P} R(S; \mathcal{M}) \quad (1)$$

In this equation, R is revenue, S is a product line comprised of P products, and \mathcal{M} is a set of choice models that account for parametric and structural uncertainty. Parametric uncertainty is considered for both the hierarchical Bayes mixed logit and latent class multinomial logit models. Structural uncertainty is represented in the latent class model by varying the number of segments.

The worst-case expected revenue for a product line is given by Equation 2, where \tilde{m} represents the choice model associated with the lowest expected per-customer revenue. Simulation results found that product line solutions that did not account for uncertainty experienced worst case losses as high as 23%. Conversely, a robust solution, using the formulations in Equations 1 and 2 could outperform a nominal solution (where it is assumed that the choice model is known precisely when the product line is optimized) by up to 14%.

$$R(S; \mathcal{M}) = \min_{\tilde{m} \in \mathcal{M}} R(S; \tilde{m}) \quad (2)$$

It is also discussed in [9] that the optimization problem given by Equation 1 may be overly conservative; that is, the perceived impact of uncertainty is dependent on how closely the uncertainty set \mathcal{M} describes the consumer population. A constrained optimization problem formulation is presented that maximizes revenue using a nominal choice model while constraining worst-case revenue to a predefined amount, as in Equation 3.

$$\begin{aligned} \max_{S \subseteq \{1, \dots, N\}: |S|=P} R(S; m) \\ \text{subject to: } R(S; \mathcal{M}) \geq \underline{R} \end{aligned} \quad (3)$$

This formulation requires accommodating a constraint violation in the fitness function (making the optimization more challenging) and an “educated” approximation of the threshold for worst-case revenue, \underline{R} . While a weighted-sum objective is also discussed that trades the performance of nominal and worst-case solutions, weighted-sum formulations have noted limitations [21].

Rather than pursue a weighted sum strategy, this paper introduces a multiobjective problem formulation that provides computational savings (in that the Pareto efficient frontier is found in a single optimization run) while allowing the tradeoff between nominal and worst-case revenue to be explored. Additionally, the problem formulations listed in Equations 1–3 model the impact of parameter (and/or structural) choice model uncertainty for the entire line. Changes in revenue represent consumers moving from a product offered by the firm to one that is offered by a competitor (or vice versa).

These works do not consider the ramifications of a choice model that reflects attributes of a product that will be physically manufactured, distributed, and sold. While revenue of the product line is still a driving business objective, the distribution of sales within the product line will dictate the allocation of resources to inventory and manufacturing. It would be expected that uncertainty in the choice model would cause variation in choice amongst the products within the line. A firm looking for a robust product design strategy would also want to minimize the

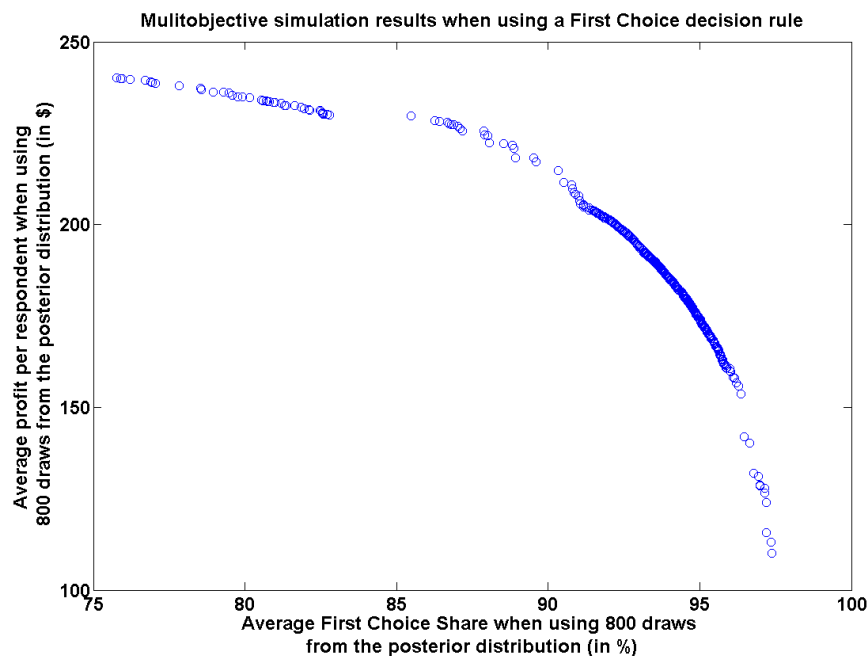
variation in individual product share. Therefore, as part of this work a third objective is introduced that minimizes *the variation in choice amongst the products within the product line*.

Exploring solution performance variation when using samples of the posterior distribution.

Previous work presented at the Sawtooth Software conference discussed the advantages of using a multiobjective optimization formulation for product line design problems. Often, however, the simulations driving the optimization use the mean of the posterior distribution from a hierarchical Bayes mixed logit model. This raises a concern when thinking about uncertainty in product line design problems—while the mean of the posterior distribution provides a Pareto efficient frontier, as shown in Figure 1, how large is the “scatter” around each Pareto point when plotting a subset of the draws used to arrive at the posterior mean?

This exploration began by using a multiobjective genetic algorithm (MOGA) to solve a product line design problem with two objectives. Part-worth estimates for 205 respondents were found using Sawtooth Software’s CBC/HB module [22]. 800 draws of the lower-level posterior distribution were saved (e.g., 800 draws per respondent) and then used in a market simulator. The modeled objectives were maximizing the average of first choice share (in percent) and the average of profit per respondent obtained by the line (in dollars). It was confirmed that the average of the part-worths across the 800 draws matched the reported mean of the posterior distribution. A first choice rule was used, and the design problem consisted of 5 products, each with 7 configuration variables. The price for each product was set as continuous variables bounded between a lower and upper bound, resulting in a mixed-integer problem formulation of 2 objectives and 40 total design variables.

Figure 1. Pareto frontier obtained when using the average of 800 draws per respondent of a HB-ML model. A first choice rule was used to model respondent choice.



The genetic search converged within 300 generations, and 422 non-dominated solutions were identified. Product configurations and prices were recorded for each solution. From these 422 solutions there were 78 unique product line configuration combinations. The remaining solutions were non-unique in that they were priced differently from another product line with a similar content configuration. Four of these solutions were then chosen for further analysis. Two of the solutions were chosen near the extremities of the identified Pareto frontier. The configuration and prices associated with these solutions are shown in Tables 1 and 2. The other two were selected near the “knee” of the Pareto frontier.

Multiple product configurations are needed because customer preferences are heterogeneous and competition exists from the outside good and competitor products that were included in the market simulator. When maximizing a share objective, as shown in Table 1, an optimization algorithm will often drive product prices to their lower bound (for this problem, \$52). Because a first choice rule is used, the optimal price for all products does not need to be at this value. Rather, they need to be at a price that does not trigger the change in binary outcome (chosen/not-chosen).

Table 1. Product configuration and pricing when maximizing the objective of average First Choice Share.

Product	Att1	Att2	Att3	Att4	Att5	Att6	Att7	Price	Avg. First Choice Share captured by each product
P1	8	8	3	6	8	6	3	\$52	28.75%
P2	8	5	3	4	4	3	4	\$52	36.55%
P3	8	8	3	4	6	8	4	\$180.50	19.66%
P4	8	5	3	4	6	3	3	\$209.03	7.33%
P5	5	8	3	6	8	1	3	\$499.07	5.07%

Maximizing the average profit per respondent requires increasing the average price of the product line. As shown in Table 2, the low-end products found in Table 1 have been replaced with products priced around \$200. These products will capture a majority of the share within the line, but the solution trades a reduction in market share for increased profit.

Table 2. Product configuration and pricing when maximizing the objective of average profit per respondent.

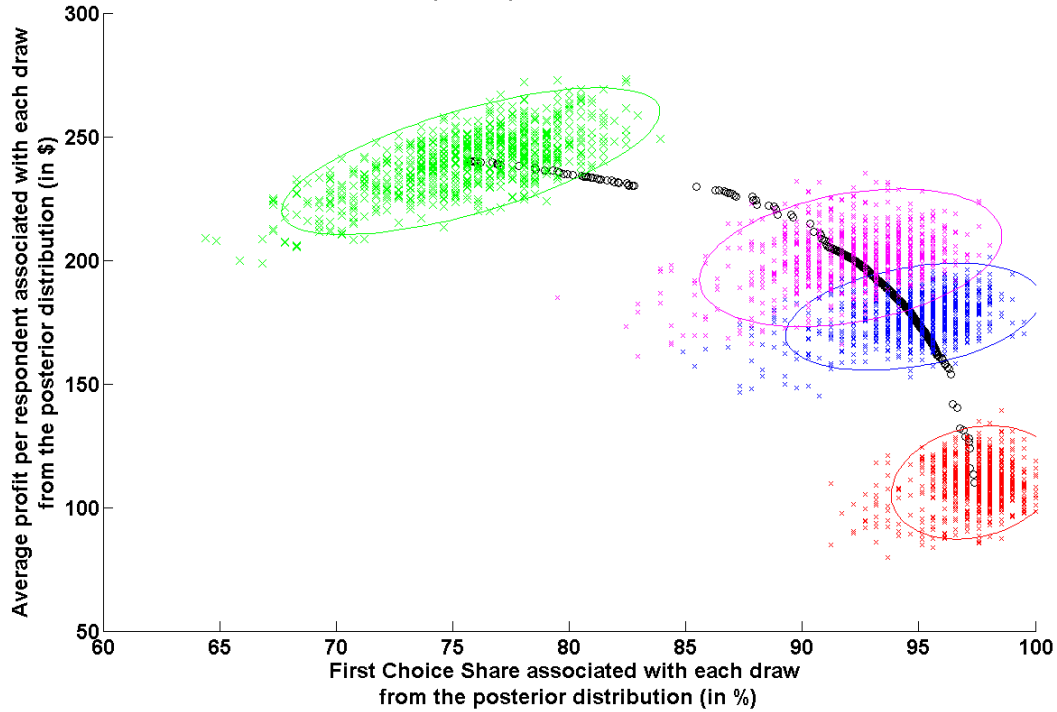
Product	Att1	Att2	Att3	Att4	Att5	Att6	Att7	Price	Avg. First Choice Share captured by each product
P1	8	5	3	4	2	3	4	\$202.03	22.38%
P2	8	5	3	4	2	3	3	\$203.13	21.29%
P3	8	8	3	4	5	8	4	\$434.10	13.59%
P4	8	8	3	6	8	6	3	\$458.87	8.73%
P5	5	8	3	6	8	1	4	\$512.88	9.78%

For the four solutions identified, the performance (average First Choice Share, average profit per respondent) of all 800 draws is shown in Figure 2. A 95% confidence interval ellipse is also shown for each solution. Immediate observations from this figure include that the confidence ellipse for the maximum share solution (red) has a smaller major axis than the maximum profit solution (green). The smaller major axis for the maximum share solution is likely due to the lower price of the first two products. These products do not make money, rather they provide a buffer against part-worth variations—they capture a large amount of share and the small

variations associated with the draws do not change this outcome. Conversely, the minor axis for the maximum share solution is larger than the maximum profit solution.

Figure 2. A plot demonstrating the scatter of performance values associated with the 800 draws per respondent from four product line solutions.

Outcomes for individual draws for 4 locations on the Pareto frontier when using a First Choice decision rule. Ellipses represent a 95% interval.



The solutions near the knee of the Pareto frontier were chosen because they highlight another challenge presented by uncertainty analysis. Dominance between two designs is no longer determined using a single set of performance values $\{F1, F2, \dots, Fn\}$. Rather, an overlap of confidence intervals opens the possibility for strict dominance to not be maintained. While there has been work on multiobjective optimization algorithms capable of handling uncertainty [23], further studies are needed so that the ramifications for market-driven product design can be better understood.

Creating a multiobjective problem formulation for robust product line design.

Parameter uncertainty was shown to have an effect when considering problem formulations driven by two different business objectives; variations of maximizing share and maximizing profit. As previously discussed, Bertsimas and Misisic proposed a problem formulation for robust problem line optimization that maximizes the worst case expected revenue given an uncertainty set (Equations 1 and 2). They also discuss how this problem statement could be reformulated to maximize revenue around a nominal choice model subject to maintaining a revenue that is no lower than some predefined amount (Equation 3). Yet, it is challenging to define this amount a priori, and constraints increase the difficulty of creating an effective fitness function.

Such challenges can be overcome by reformulating the constraint-based problem described in Equation 3 as a multiobjective optimization problem. This formulation is constructed around a

nominal model (m). For the purpose of this exploration, the mean of the posterior distribution was used as the nominal model because it corresponds to the part-worth values used in previous market simulator implementations. The uncertainty set (\mathcal{M}) consisted of the mean of the posterior distribution and the 800 draws (a subset of the total draws) saved from estimating the posterior distribution. The lower bound on product price was also redefined. Price was set at 125% of product cost, plus a constant value that was consistent across all products offered by the manufacturer. The objectives for the optimization were defined as maximizing the revenue per respondent (in dollars) and maximizing the worst case revenue across the uncertainty set, per respondent (in dollars). It should be noted that this is different than the formulation proposed by Bertsimas and Misic who use worst case expected revenue. This formulation for objective F2 is heavily weighted toward the worst case scenario, and the full formulation is given by Equation 4.

<p>Nominal model = Mean of the posterior distribution</p> <p>Uncertainty set = 800 draws (per respondent) from the, and the mean of the, posterior distribution</p> <p>Product price = $1.25 \times \text{Product cost} + \\52</p> <p>Number of products = 5 (with 7 configuration variables each)</p> <p>Use a multiobjective genetic algorithm (MOGA) to solve:</p> <p style="padding-left: 20px;">Maximize: F_1 = Revenue per respondent using the nominal model (in \$)</p> <p style="padding-left: 20px;">Maximize: F_2 = Worst case revenue from uncertainty set, per respondent (in \$)</p>	(4)
---	-----

The problem statement given by Equation 4 was optimized using a multiobjective genetic algorithm. Because product price was now a function of configuration cost, the number of unique solutions decreased. As shown in Figure 3, 8 unique product line configurations were identified as Pareto optimal points. Solutions in the upper right corner of the graph are preferred, as they maximize both revenue in the nominal model and the worst case revenue from the uncertainty set. Numerical results for these 8 solutions are presented in Table 3. In this table the maximum revenue per respondent is presented when using the mean of the posterior distribution (the nominal model). The worst case revenue, mean revenue, and the largest revenue, recorded from the 800 draws of the posterior distribution are also presented for each solution.

Figure 3. Pareto frontier when maximizing revenue per respondent under the nominal model and maximizing worst case revenue from the uncertainty set.

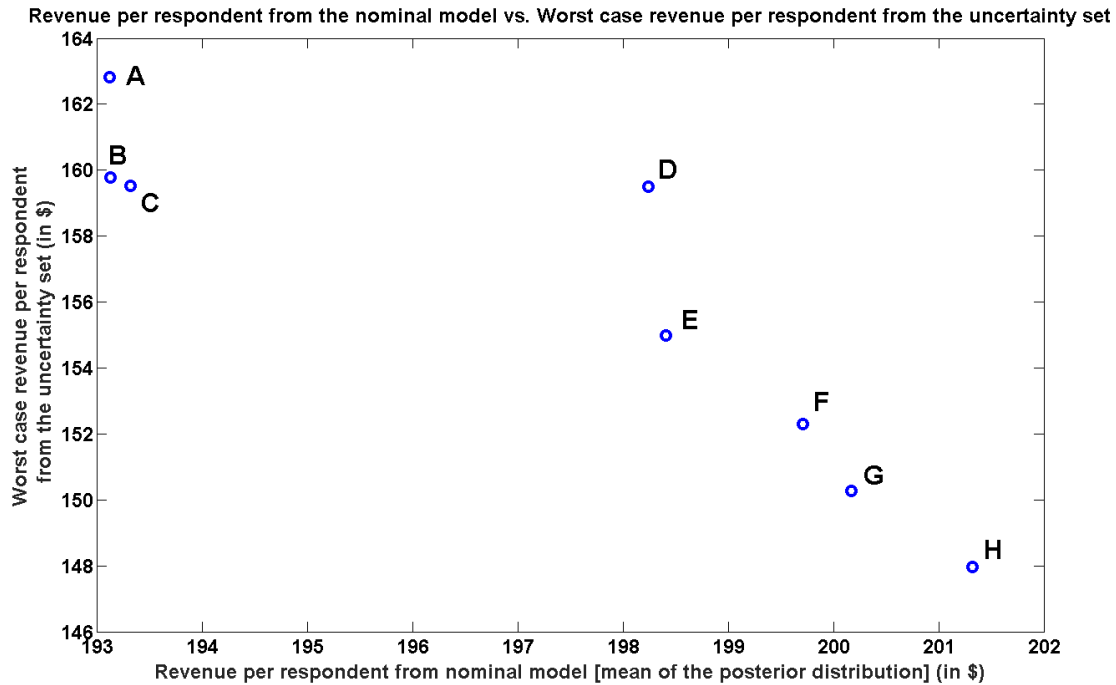


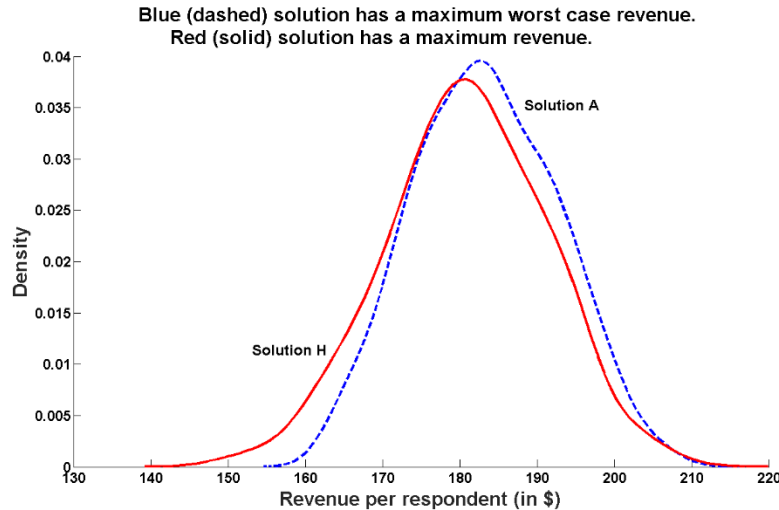
Table 3. Revenue values for the eight Pareto frontier solutions determined using the nominal model and the uncertainty set.

Solution	Maximum revenue per respondent calculated from mean of posterior distribution	Revenue per respondent calculated from samples of the posterior distribution		
		Minimum (worst case)	Mean	Maximum
A	\$193.13	\$162.81	\$183.30	\$208.17
B	\$193.14	\$159.78	\$185.49	\$211.19
C	\$193.32	\$159.52	\$185.08	\$211.30
D	\$198.24	\$159.49	\$186.50	\$211.44
E	\$198.41	\$154.99	\$179.41	\$203.64
F	\$199.71	\$152.29	\$180.59	\$207.41
G	\$200.17	\$150.27	\$181.98	\$208.79
H	\$201.32	\$147.96	\$180.75	\$211.15

The samples from the posterior distribution can also be used to create a probability density solution. Two of these distributions are shown in Figure 4. By moving from left to right in Figure 3, the worst case revenue decreases. Figure 4 illustrates that in the presence of parameter

uncertainty a solution designed around the criterion of maximizing worst case revenue has opportunities to outperform a solution purely designed for maximum revenue. This creates a scenario where a decision-maker must define the level of risk they are willing to adopt. Existing engineering design tools for concept selection provide insight into how such decisions can be made using utility theory and hypothetical alternatives [24].

Figure 4. Probability density plot for solutions that maximize revenue (solid) and maximize worst case revenue (dashed).



At the suggestion of Bryan Orme, the revenue over the 800 draws were examined. The concern was that an outlier would make a worst-case revenue objective too aggressive. While all revenue values were found to be within 3.5 standard deviations of the mean, the lack of outliers does not eliminate the significance of this concern. A more effective strategy for this objective may involve defining a worst-case revenue percentile the decision-maker is willing to accept. For the eight solutions found in this study, changing from worst case revenue to revenue at the 1st or 5th percentiles can cause solutions to become dominated. Here, Solution B would dominate Solution A, removing A as a Pareto point and preventing it from ever being chosen. The worst case revenue, and the revenue at the 1st and 5th percentiles, for each solution are shown in Table 4.

Table 4. Revenue values for the eight Pareto solutions when considering worst case revenue, 1st percentile revenue, and 5th percentile revenue.

Solution	Worst case revenue	1 st percentile revenue	5 th percentile revenue
A	\$162.81	\$163.96	\$168.32
B	\$159.78	\$165.25	\$170.04
C	\$159.52	\$165.26	\$169.83
D	\$159.49	\$164.60	\$171.10
E	\$154.99	\$155.84	\$164.65
F	\$152.29	\$159.27	\$165.73
G	\$150.27	\$159.00	\$164.80
H	\$147.96	\$154.43	\$163.06

Concerns about the effect of parameter uncertainty on business objectives aligned with revenue motivated further analysis. If the uncertainty set could be used as a means of

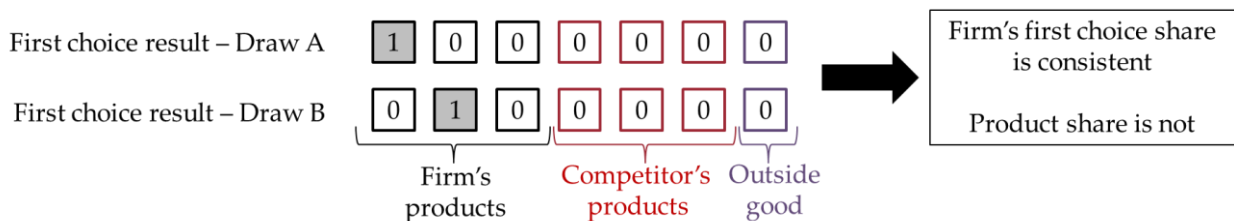
determining worst-case revenue, then the uncertainty set could also be used to explore variations of first choice share *within a product line*. Uncertainty analysis conducted this way provides a product line perspective that has not been discussed in the literature, and is discussed in the next section.

Exploring the effect of parameter uncertainty from a product share perspective.

Additional motivation for exploring the effect of parameter uncertainty from a product share perspective is shown in Figure 5. Consider a scenario where the manufacturer has decided to offer three products. These products compete in a market against three competitor products and an outside good. Now, consider a single respondent making a choice in this market. Their selection is modeled using a first choice decision rule. For a given draw from the posterior distribution (let us call it Draw A), results from the market simulator indicate that the respondent selects the first of the three products offered by the manufacturer. Since a first choice rule is being used, choice is fully assigned to a single product.

The uncertainty set discussed in the previous section was comprised of a set of draws from the posterior distribution. If another draw is considered (we will call this one Draw B), the results from the market simulator indicate that the *same respondent* has now chosen the second of the firm's three offerings. From the share perspective of a product line, nothing has changed; the effect of uncertainty would be unobservable. Yet, from a product manufacturing and component inventory perspective, the change in respondent choice is significant. While the first choice share for this respondent at the product line level remains consistent at 100%, the deviation of share within the product line is also 100% (going from the first offering to the second offering).

Figure 5. Representative example demonstrating how parameter uncertainty can be unobservable at the product line level, while having significant impact at the product share level.



This led to a question that motivated the second half of this work—how big of an issue is parameter uncertainty when making resource and production allocation decisions? As an initial exploration, the variability in product share was examined for Solution A from Figure 3. Observations for First Choice Share within the product line were taken over the 800 draws from the posterior distribution. A summary of these observations are reported in Table 5. Reported values include the First Choice Share of each product in the line from the nominal model (the mean of the posterior distribution) and the mean, standard deviation, minimum, and maximum values of first choice share distribution from the 800 draws.

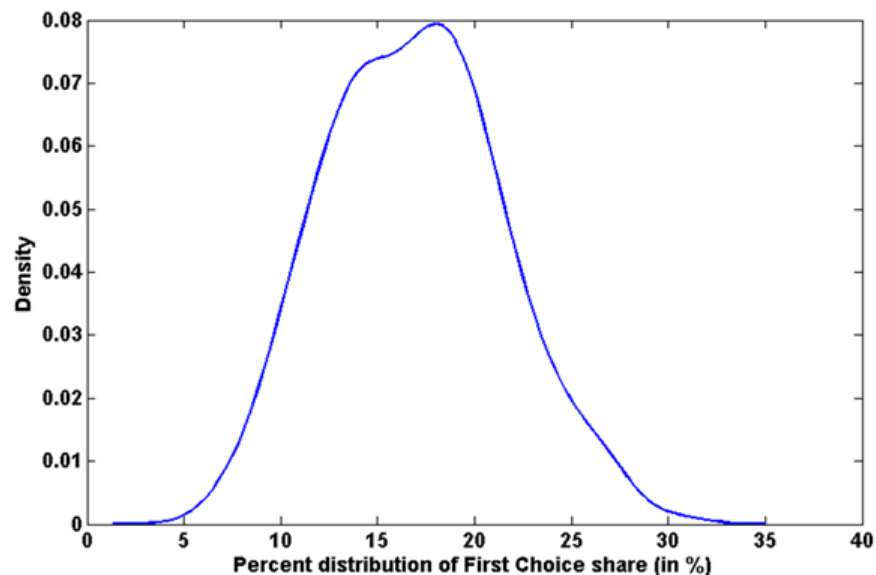
The significance of parameter uncertainty when making configuration and pricing decisions is demonstrated by the misalignment of within line share distribution between the nominal model and the uncertainty set. First, there is a difference in the mean values of First Choice Share between the nominal model and the uncertainty set. Perhaps more important is the range between minimum and maximum percent distribution of First Choice Share. The density plot for Product

1's First Choice Share is shown for the uncertainty set in Figure 6. In the next section, a problem formulation strategy designed to reduce the width of this distribution is introduced.

Table 5. Comparison of First Choice Share for products within a product line solution. Results for both the nominal model and the uncertainty set are reported.

Product within the line	Percent distribution of within line First Choice Share (in %)				
	Model: Mean of the posterior distribution	Model: Uncertainty set			
		μ	σ	Min	Max
Product 1	13.01	16.90	4.54	5.34	31.03
Product 2	9.59	12.25	3.45	2.92	21.43
Product 3	12.33	15.04	4.23	5.17	27.52
Product 4	43.15	34.77	4.97	20.97	45.65
Product 5	21.92	21.04	4.59	9.72	32.79

Figure 6. Density plot of First Choice Share distribution for Product 1 using the uncertainty set.

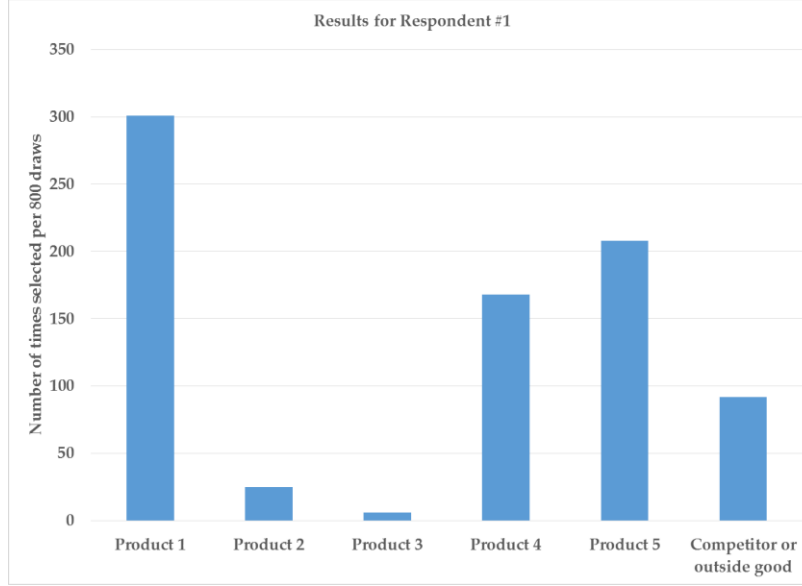


Building variation in product share from model uncertainty into the problem formulation.

We build on the results presented in the previous section by further exploring how Respondent #1's product selection changes over the 800 draws from the posterior distribution. A bar chart showing choice rule outcomes is shown in Figure 7. Using the nominal model of the mean of the posterior distribution, the choice rule results in a selection of Product 1 from the firm. For approximately 300 of the 800 draws, this choice rule result is also obtained. Less than

100 of the draws divert share from the firm to the competitor products or the outside good. Rather, over half of the draws from the posterior distribution maintain firm-level share while diverting product share, with most of the share going to Products 4 and 5.

Figure 7. First choice rule results for Respondent #1 using the uncertainty set.
Under the nominal model this respondent's first choice was Product 1.



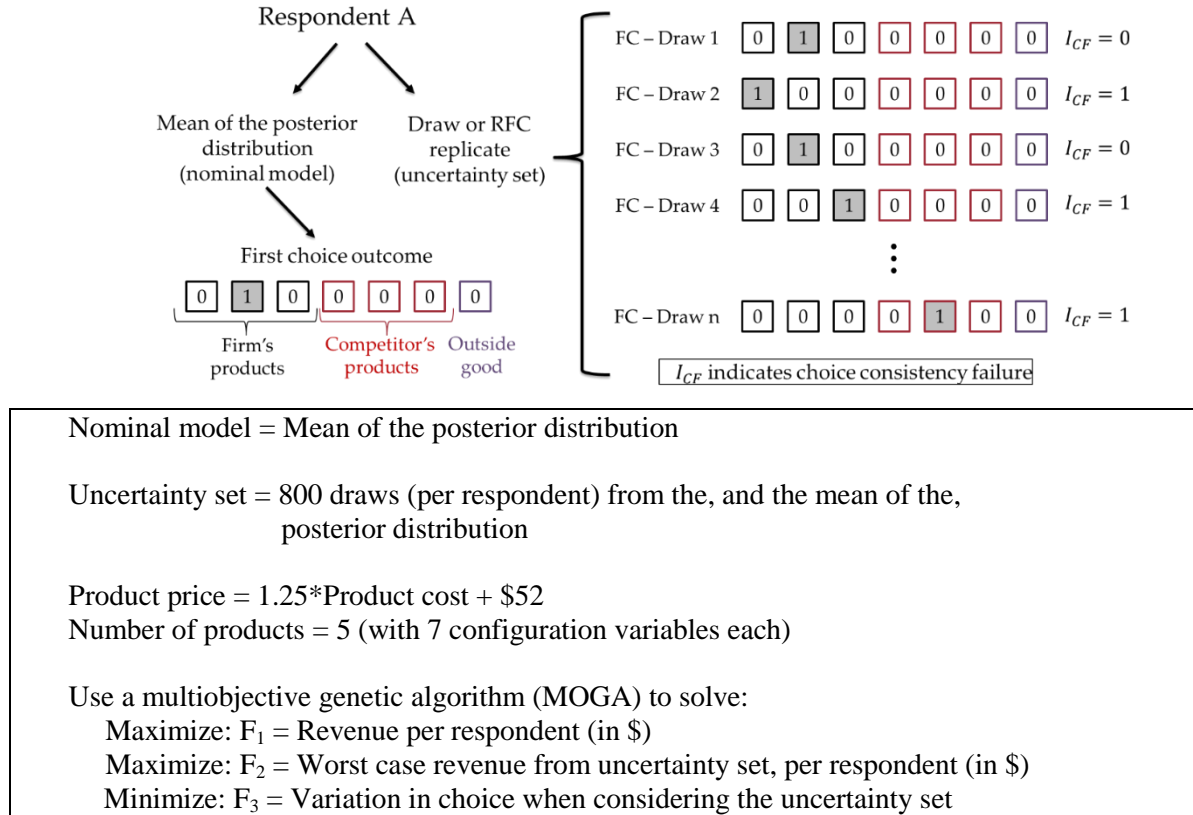
These results support the argument that a robust product line solution should be one that captures maximum market share with minimal variability, while *also* minimizing the variation in product share. A first thought was to develop a metric that quantified choice consistency so that it could be incorporated into a multiobjective problem formulation. This metric, as shown in Equation 5, calculates the average number of choice inconsistencies per respondent. Here, N is the number of respondents who chose one of the firm's products using the first choice rule and the mean of the posterior distribution. R is the number of draws (for this problem 800).

As shown in Figure 8, the mean of the posterior distribution is used as the “truth” for each respondent. The results of the first choice rule using these point-estimates are then compared against the market simulator results for all 800 draws. For cases where the choice rule outcome between the nominal model and one of the draws of the uncertainty set align, the indicator function is 0. When the outcome of the choice rule between the nominal and uncertainty model are different, the indicator function is 1.

$$ACI \equiv P[\mathbf{c} \notin \Omega_{PL,D}] = \frac{1}{N} \sum_{n=1}^N \sum_{r=1}^R I_{CF}^{n,r}(\mathbf{c}_{n,r}) \quad (5)$$

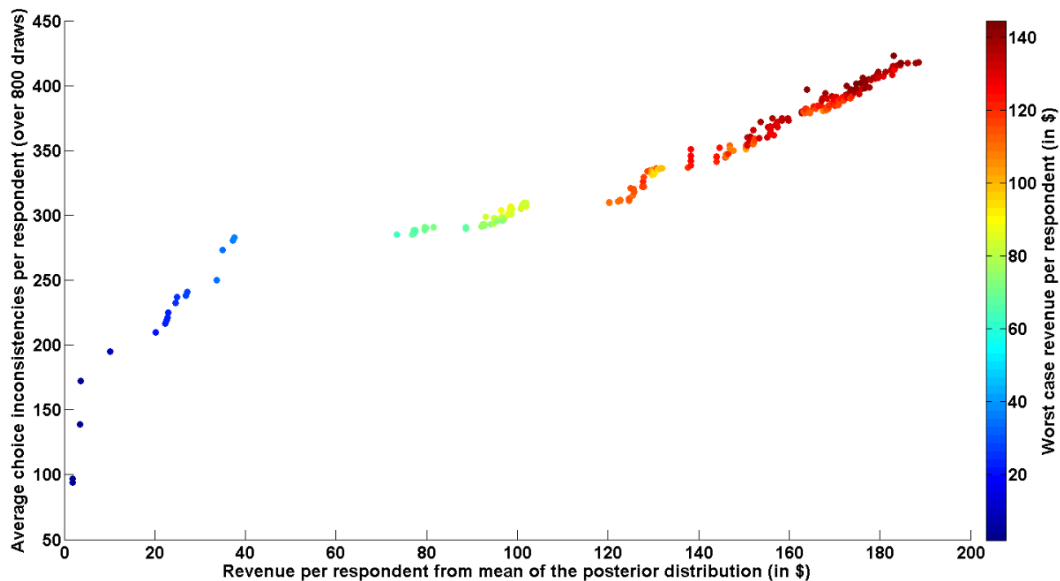
The introduction of Equation 5 allows for the formulation of a three-objective optimization problem. This problem, shown in Equation 6, builds on the previous formulation with the additional goal of minimizing the average number of choice inconsistencies per respondent. 318 unique solutions were found using this optimization problem. As might be expected, the design solutions found to be most robust from choice inconsistencies performed poorly on revenue objectives. A two-dimensional scatterplot is shown in Figure 9 that illustrates the tradeoff between revenue per respondent and average choice inconsistencies per respondent. Solutions that reduce choice inconsistencies lead to product line solutions that generate minimal revenue.

Figure 8. Representation of choice consistency failure when considering multiple draws. The first choice decision under the nominal model is used as the reference.



(6)

Figure 9. Scatterplot of the optimal designs generated for the three-objective problem formulation. As average choice inconsistencies decrease, so too does the revenue generated by the line.



Reformulating the optimization problem for market-level behavior.

The solutions shown in Figure 9 are based on the concept that choice inconsistencies should be minimized for each respondent. Yet, this formulation may not reflect the actual goal. The thought of reformulating the third objective was inspired by Chris Chapman's paper at the 2013 Sawtooth Software conference [25]. In this paper, Chris discussed how the results of a market simulator were not intended to focus on the behavior of an individual respondent, but the overall response of the market as a whole.

This led to a realization: choice inconsistencies at the respondent level could cancel each other out, but this was not accounted for in Equation 5. Rather, this outcome was being penalized twice. The choice inconsistency formulation shown in Equation 5 was then replaced with a variation in First Choice Share (FCS) calculation shown in Equation 7. Here, n represents the number of products being developed by the manufacturer. The first choice share is calculated for each product using the nominal model. The difference between the FCS from the nominal model and the mean FCS obtained from the uncertainty set is then determined. This result is squared and multiplied by a weighting factor w_i . Weighting factors are bounded between 0 and 1, and the sum of the weighting factors must equal 1. While the weighting factors for most problems may be equal, the weight in FCS deviation can be increased for a particular product when it has configuration parameters specific to it. For example, a single product in the line may use a unique engine type, or a particular material, that could not be used on other products in the line if manufacturing numbers are adjusted. Equation 7 is then used in the reformulated three-objective optimization problem shown in Equation 8.

$$\sqrt{\sum_{i=1}^n w_i (FCS(i)_{nominal} - \mu(FCS(i))_{uncertainty\ set})^2} \quad (7)$$

$$\sum w_i = 1$$

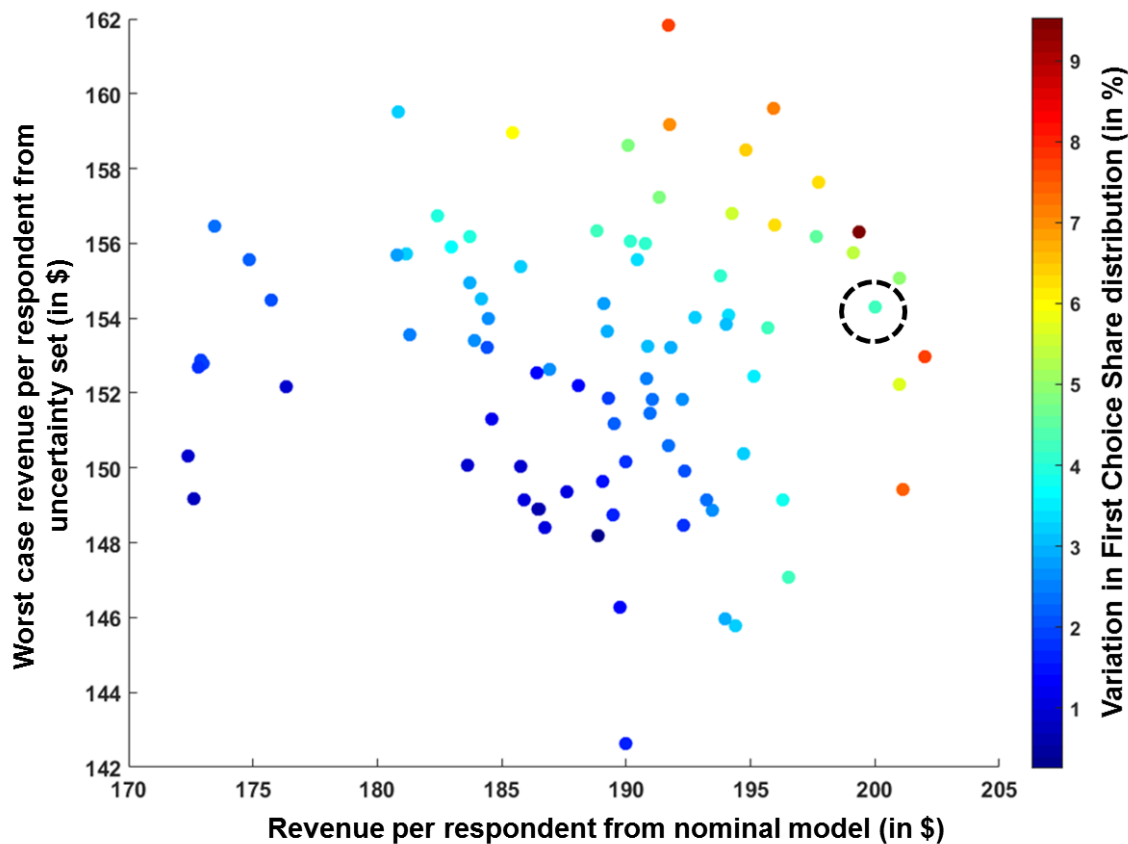
$$0 \leq w_i \leq 1$$

<p>Nominal model = Mean of the posterior distribution</p> <p>Uncertainty set = 800 draws (per respondent) from the, and the mean of the, posterior distribution</p> <p>Product price = 1.25*Product cost + \$52</p> <p>Number of products = 5 (with 7 configuration variables each)</p> <p>Use a multiobjective genetic algorithm (MOGA) to solve:</p> <p style="padding-left: 20px;">Maximize: F_1 = Revenue per respondent (in \$)</p> <p style="padding-left: 20px;">Maximize: F_2 = Worst case revenue from uncertainty set, per respondent (in \$)</p> <p style="padding-left: 20px;">Minimize: F_3 = Variation in First Choice Share distribution (in %)</p>	(8)
---	-----

Solving this optimization problem allows for the simultaneous consideration of business and manufacturing tradeoffs in the presence of parameter uncertainty. As shown in Figure 10, a decision-maker can identify that the product line solution with the maximum worst case revenue also has one of the highest variations in first choice share distribution. Multi-attribute decision making tools can be used when selecting a final solution from this set of non-dominated product

lines. A solution has been identified in Figure 10 that reduces the variation in first choice share distributions to under 4% for the product line that lies near the efficient frontier for the tradeoff between average revenue per respondent (from the nominal model) and the average worst case revenue per respondent (from the uncertainty set).

Figure 10. Scatterplot of the reformulated three-objective problem.
Here, the colorbar represents the variation in First Choice Share distribution of products within the line.



The configurations for this product line are shown in Table 6. Some commonality is observed in this solution—all products use configuration level 3 for the third attribute, only levels 5 and 8 are used for the second attribute, etc. Commonality reduces the concern associated with variation in first choice share caused by parameter uncertainty. However, for the sixth product attribute, four different configuration levels compose the product line solution. As part of future work, the weighting term in Equation 7 could be scaled as a function of the number of unique components used in a particular product. This would enforce greater consistency for those components that must be purchased for only a single product. Developing a greater understanding of how variation in first choice share distribution impacts inventory and supply chain decisions could reduce negative business outcomes caused by parameter uncertainty.

Table 6. Product line configuration profile for the solution identified in Figure 10.

Product	Att1	Att2	Att3	Att4	Att5	Att6	Att7
P1	4	8	3	6	7	1	4
P2	5	5	3	4	2	3	4
P3	8	5	3	6	5	4	3
P4	8	8	3	3	2	3	4
P5	8	8	3	4	5	8	3

CONCLUSIONS AND FUTURE WORK

Market simulators created from estimates of customer preference are powerful tools for exploring market response to new product offerings. When heterogeneity is represented using a hierarchical Bayes mixed logit model, the most basic market simulators will use part-worth values for each respondent associated with the mean of the lower-level posterior distribution. This choice is made because it reduces computational complexity, allowing for faster simulations and reduced cost when optimizing a product line. However, a failure to account for uncertainty can undermine these computational advantages and result in product configuration and pricing decisions that forfeit value.

The simulations presented in this paper demonstrate the importance of accounting for uncertainty when conducting market simulations. Parameter uncertainty was addressed by considering 800 draws saved from the lower-level posterior distribution of a hierarchical Bayes mixed logit model. For a multiobjective optimization problem of share versus profit, four locations of the Pareto frontier were explored. As more weight is placed on the profit objective, the major axis of the confidence ellipse grew (meaning more scatter in the predicted share of the product line over the 800 draws). Conversely, the minor axis of the confidence ellipse shrank, leading to less scatter on the profit objective. Comments have been made at previous Sawtooth Software conferences about the possibility of an overstated variance for uncertainty when simulating from draws obtained from the lower-level posterior distribution. Simulating from the upper-level model may provide a more accurate uncertainty representation, though lower-level models have been shown to reflect respondent heterogeneity for product line problems. Exploring how information from both the upper- and lower-level models can be used is an opportunity for future work.

Attention then turned to an optimization problem that maximized the revenue generated by a product line. Here, the mean of the posterior distribution was used as the basis for a nominal model. This mean value from the lower-level HB model and the 800 draws from the posterior distribution were combined to create an uncertainty set of models. A multiobjective optimization problem was formulated that maximized revenue under the nominal model while simultaneously maximizing worst case revenue from the uncertainty set. By looking at worst case revenue from an entire set of models, an effective “lower bound” for revenue could be determined for each product line solution.

There are interesting challenges raised by this problem formulation. The problem formulation given by Equation 4 considers a worst case scenario. Outliers may drive the optimization result, and the decision to define a percentile threshold may be more effective. As the paper’s discussant at the conference, Mark Beltramo noted that maximizing worst-case revenue amounts to maximizing revenue from a small quantile of the distribution, corresponding to a decision-maker

that is extremely risk averse. Because optimizing the nominal model provided a biased estimate for maximum revenue compared to the average revenue over the uncertainty set, he continued by proposing a problem formulation where the first objective maximized average revenue per respondent over the uncertainty set (expected return), while the second objective maximized average revenue per respondent minus the k th quantile over the draws (risk). Since robust product line design is similar in concept to balancing risk and return in a stock portfolio [26], he suggested that the 5th percentile may be considered for k to align with common practice in finance. Finally, the composition of the uncertainty set could be further explored. All draws that were saved from the posterior distribution are considered to have equal value, though it may be true that some are closer to resembling true market behavior.

A significant contribution of this paper is the introduction of a third objective that aligns with the resource and production allocation decisions discussed by Bertsimas and Misic. The introduction of a third objective began with a measure of choice inconsistency within a product line that was measured at the respondent level. However, individual respondent choice may not be the appropriate concern. This third objective was reformulated to minimize the variation in First Choice Share for each product in the line. The two measures used were the within-line First Choice Share from the nominal model and the average within-line First Choice Share from the uncertainty set. By adding a third objective, solutions can be found that balance revenue uncertainty at the product line level while providing insight into the extent that the choice of individual products varies. This information is significant because it can be used to identify how parameter uncertainty within the market simulator might impact component ordering and production decisions that need to be made by a manufacturer. The discussant noted that expected holding costs increase with variance of demand, and a formulation of the third objective that minimized a weighted sum of the variances of the individual product choice shares could be used. Such considerations have been unexplored using market simulators, and this formulation presents significant opportunities for the use of quantitative market research models when designing products that are physically produced.

While some opportunities for future work have already been discussed, it is important to note that the results presented in this paper only consider model parameter uncertainty. Structural uncertainty has been discussed in the literature, though it was not addressed in this paper. Further, the incorporation of uncertainty into market simulators can include uncertainty related to the product attributes used by the firm, component costs, and the attributes and prices of competitor products. Incorporating these additional uncertainties will increase the computational expense of a simulation. Yet, the potential hazard of ignoring uncertainty in market simulators can lead to configuration and pricing decisions that forfeit value and can result in resource allocation decisions that cannot be easily reversed or corrected.

ACKNOWLEDGEMENTS

I would like to thank Bryan Orme for his insightful questions and comments during presentation preparation. I would also like to thank Mark Beltramo for the discussion he provided at the Sawtooth Software conference and for his insights about future work in this area.



Scott Ferguson

REFERENCES

- [1] Chapman, C., and Alford, J., 2011, "Product portfolio evaluation using choice modeling and genetic algorithms," Proc. 2010 Sawtooth Softw. Conf.
- [2] Turner, C., Foster, G., Ferguson, S., Donndelinger, J., and Beltramo, M., 2012, "Creating targeted initial populations for genetic product searches," 2012 Sawtooth Softw. Conf.
- [3] Foster, G., Ferguson, S., and Donndelinger, J., 2014, "Demonstrating the Need and Value of a Multiobjective Product Search," 2013 Sawtooth Software Conference, October 14–18, Dana Point, CA.
- [4] Ferguson, S.M., 2015, "Climbing the Content Ladder: How Product Platforms and Commonality Metrics Lead to Intuitive Product Strategies," 2015 Sawtooth Software Conference, Orlando, FL.
- [5] Tsafarakis, S., Marinakis, Y., and Matsatsinis, N., 2011, "Particle swarm optimization for optimal product line design," Int. J. Res. Mark., 28(1), pp. 13–22.
- [6] Michalek, J.J., Ceryan, O., Papalambros, P.Y., and Koren, Y., 2006, "Balancing Marketing and Manufacturing Objectives in Product Line Design," J. Mech. Des., 128(6), p. 1196.
- [7] Wang, X.F., Camm, J.D., and Curry, D.J., 2009, "A branch-and-price approach to the share-of-choice product line design problem," Manage. Sci., 55(10), pp. 1718–1728.
- [8] Belloni, A., Freund, R., Selove, M., and Simester, D., 2008, "Optimizing Product Line Designs: Efficient Methods and Comparisons," Manage. Sci., 54(9), pp. 1544–1552.
- [9] Bertsimas, D., and Mišić, V.V., 2017, "Robust Product Line Design," Oper. Res., 65(1).
- [10] Gilbride, T.J., and Lenk, P.J., 2010, "Posterior Predictive Model Checking: An Application to Multivariate Normal Heterogeneity," J. Mark. Res., 47(5), pp. 896–909.
- [11] Abramson, C., Andrews, R.L., Currim, I.S., and Jones, M., 2000, "Parameter Bias from Unobserved Effects in the Multinomial Logit Model of Consumer Choice," J. Mark. Res., 37(4), pp. 410–426.
- [12] Montgomery, A.L., and Bradlow, E.T., 1999, "Why Analyst Overconfidence About the Functional Form of Demand Models Can Lead to Overpricing," Mark. Sci., 18(4), pp. 569–583.

- [13] Orme, B.K., and Huber, J., 2000, "Improving the value of conjoint simulations," *Mark. Res.*, 12(4), pp. 12–20.
- [14] Camm, J.D., Cochran, J.J., Curry, D.J., and Kannan, S., 2006, "Conjoint Optimization: An Exact Branch-and-Bound Algorithm for the Share-of-Choice Problem," *Manage. Sci.*, 52(3), pp. 435–447.
- [15] Kleywegt, A.J., Shapiro, A., and Homem-de-Mello, T., 2002, "The Sample Average Approximation Method for Stochastic Discrete Optimization," *SIAM J. Optim.*, 12(2), pp. 479–502.
- [16] Wang, X. (Jocelyn), and Curry, D.J., 2012, "A robust approach to the share-of-choice product design problem," *Omega*, 40(6), pp. 818–826.
- [17] Luo, L., Kannan, P.K., Besharati, B., and Azarm, S., 2005, "Design of Robust New Products under Variability: Marketing Meets Design*," *J. Prod. Innov. Manag.*, 22(2), pp. 177–192.
- [18] Besharati, B., Luo, L., Azarm, S., and Kannan, P.K., 2006, "Multi-Objective Single Product Robust Optimization: An Integrated Design and Marketing Approach," *J. Mech. Des.*, 128(4), p. 884.
- [19] Resende, C.B., Grace Heckmann, C., and Michalek, J.J., 2012, "Robust Design for Profit Maximization With Aversion to Downside Risk From Parametric Uncertainty in Consumer Choice Models," *J. Mech. Des.*, 134(10), p. 100901.
- [20] Ver Hoef, J.M., 2012, "Who Invented the Delta Method?," *Am. Stat.*, 66(2), pp. 124–127.
- [21] Das, I., and Dennis, J.E., 1997, "A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems," *Struct. Optim.*, 14(1), pp. 63–69.
- [22] Sawtooth Software inc., 2014, "Sawtooth Software CBC/HB 5.5.3."
- [23] Chapman, J.L., Lu, L., and Anderson-Cook, C.M., 2014, "Incorporating response variability and estimation uncertainty into Pareto front optimization," *Comput. Ind. Eng.*, 76, pp. 253–267.
- [24] See, T.-K., Gurnani, A., and Lewis, K., 2004, "Multi-Attribute Decision Making Using Hypothetical Equivalents and Inequivalents," *J. Mech. Des.*, 126(6), p. 950.
- [25] Chapman, C., 2013, "9 Things Clients Get Wrong about Conjoint Analysis," *Proceedings of the 2013 Sawtooth Software Confernece*, B. Orme, ed., Dana Point, CA.
- [26] Markowitz, H.M., 1952, "Portfolio selection," *J. Finance*, 7(60), pp. 77–91.

PROPERTIES OF DIRECT UTILITY MODELS FOR VOLUMETRIC CONJOINT

JAKE LEE
QUANTUM STRATEGY, INC

ABSTRACT

Volumetric conjoint models are an exciting, new area for choice modeling practitioners. The new models are based on established economic theory and don't require duct tape. The models are very new and still need investigation to understand the circumstances when they work well and when adjustments need to be made.

Direct utility models are more appropriate when you'd expect consumers to pick multiple options to maximize their utility. The model accounts for 2 ideas that may be new to the choice modeling community. Specifically, the model has parameters of budget constraint and satiation, to give a more complete understanding of the consumer choice process.









A CASE STUDY—MACARONI AND CHEESE

In 2016, macaroni and cheese sales were near \$1.5B. With over 1,000 SKUs in total, the top 40 SKUs represent 72% of total category revenue. And a small number of brands made up the majority of sales.

We conducted an in-house study (no client) of this category to take a closer look at modeling options and impacts on managerial decision making. Here is an example choice task:

Think about your shopping for the next month. How many of each of the following would you buy? (Type in any number)

If you wouldn't buy any of a particular option, please type in zero before continuing.

 <ul style="list-style-type: none">• Cheddar Shells with Powder Cheese Packet• Microwave• 10 grams of protein  <p>Single pack 7.25 ounce box \$2.50</p> <input type="text" value="0"/>	 <ul style="list-style-type: none">• White Cheddar Elbow Macaroni with Powder Cheese Packet• Stove Top or Microwave  <p>Single pack 2 ounce cup \$1.00</p> <input type="text" value="0"/>	 <ul style="list-style-type: none">• Sharp Cheddar Fun Shapes with Liquid Cheese Packet• Microwave• 10 grams of protein  <p>12 pack 7.25 ounce box \$9.00</p> <input type="text" value="0"/>	 <ul style="list-style-type: none">• Wisconsin Cheddar Spiral with Liquid Cheese Packet• Microwave• Organic  <p>18 pack 2 ounce cup \$31.50</p> <input type="text" value="0"/>
--	--	---	--

TOTAL: \$0

We collected 1,000 responses from consumers who had purchased any macaroni and cheese in the last 3 months. Each respondent provided purchase estimates for 8 scenarios. The average survey completion time came in just under 6 minutes. In addition to the choice exercise we asked just a few behavioral, attitudinal and demographic questions.

Just like in the grocery store, respondents were free to choose any number of boxes or cups. They also could choose to not purchase any by selecting zero for each option. For convenience, the total cart amount was dynamically calculated for the respondent to see.

The design elements were chosen after reviewing syndicated sales data for the category. They were chosen to accommodate most of the top 40 highest revenue products to later be put into the simulator. For fun, we included a few features that were not currently in the top products, but could be used as potential innovation in the category.

In the chart below, each row represents an attribute with its various levels displayed from left to right.

att/lev	1	2	3	4	5	6	7	8
Brand	Velveeta	Kraft	Annie's	Cracker Barrel	Pasta Roni	Private Label	Pepperidge Farm	Horizon
Container	7.25 ounce box	2 ounce cup						
Pasta Type	Elbow Macaroni	Shells	Spiral	Fun Shapes				
Flavor	Cheddar	3 Cheese	White Cheddar	Sharp Cheddar	Creamy	Wisconsin Cheddar		
Cheese Packet	powder Cheese Packet	Liquid Cheese Packet						
Cooking	Stove Top	Microwave	Stove Top or Microwave					
Claim		Organic	Gluten Free	Low Calories	10 grams of protein			
Pack Size	Single pack	4 pack	8 pack	12 pack	18 pack			
Box Price	\$0.75	\$1.00	\$1.50	\$2.00	\$2.50	\$3.00	\$3.50	
Cup Price	\$0.50	\$0.75	\$1.00	\$1.25	\$1.50	\$1.75	\$2.00	

We included separate price ranges for boxes and cups to be consistent with the market values. The final price shown to respondents was calculated by multiplying the number of units in the package by the price level assigned to that alternative.

A randomly generated experimental design was used to tee up product configurations for the respondents (more on experimental design later).

After cleaning the data for consumers with zero demand and bad responses, the total sample size was 958.

CONCERNS ABOUT RESPONDENT QUALITY—PROPOSED INTERACTIVITY

Experienced researchers will have the same concern moments after seeing an example choice task with a volumetric response: What if respondents type in unreasonably large numbers? Won't that have a large/undue impact on the model results?

The answer is yes—unreasonably large responses can have a very dramatic impact on the results. Beyond cleaning respondents who give unbelievable responses, there are a few interviewing strategies that you can use to keep the responses closer to reality.

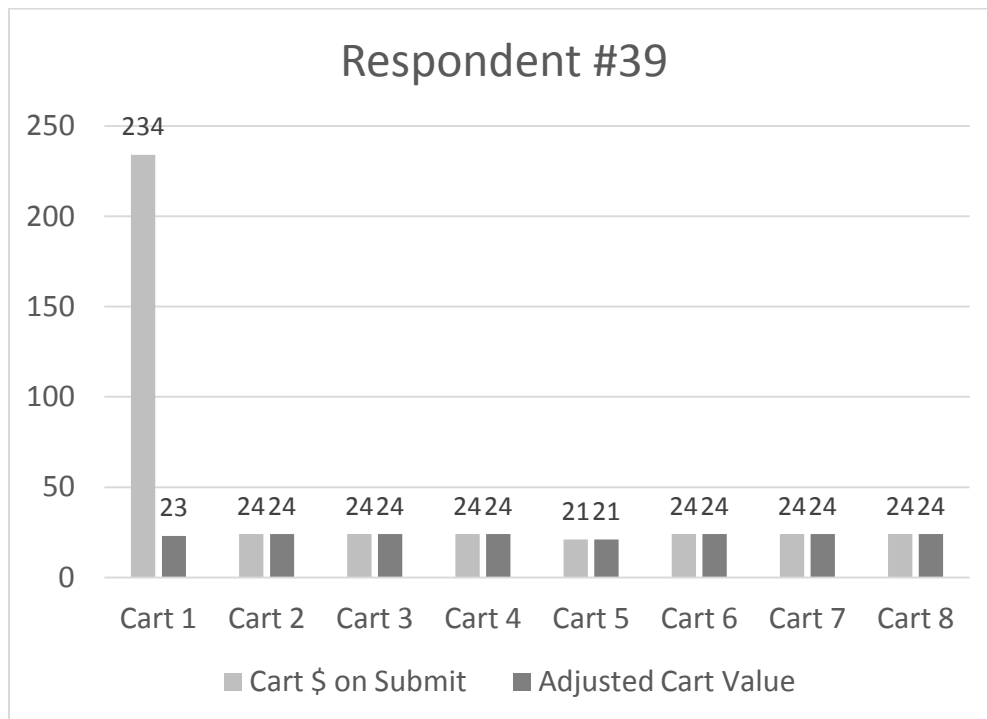
Our proposed interactive interview process has 4 steps:

1. Ask about past purchase behavior.
2. Ask for a reasonable budget amount.
3. Identify a threshold for what you consider “high” spend.
4. If the total cart amount for any task exceeds the value from 2 or 3, provide an opportunity for them to revise their cart.

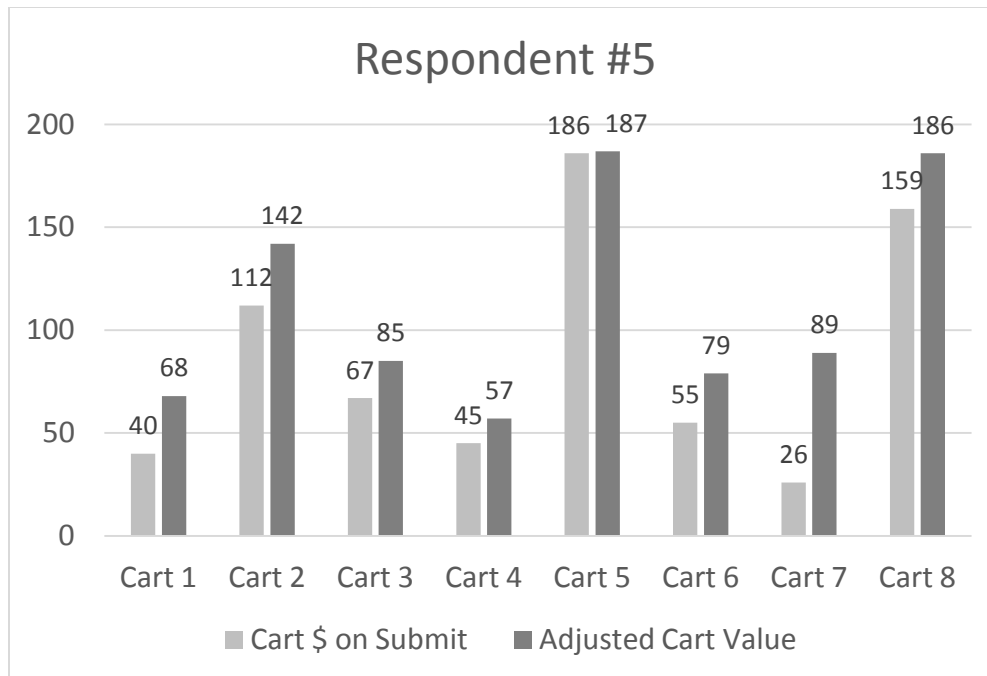
The interactive approach helped us to identify outliers or bad respondents. It also provided a mechanism for respondents to catch their own typos.

For example, this respondent filled their cart with \$234 worth of macaroni and cheese on their first task. After being prompted with an option to revise their cart, the dollar amount was reduced to \$23.

It looks like the first submission had a typo that otherwise would have led to having all of the respondent's data tossed.



Another respondent had some very weird response patterns. Each cart had high enough \$ amount to trigger a prompt. After each prompt the respondent added more products to their cart. This behavior was exactly what we'd expect from a bot and they were removed from the final data.



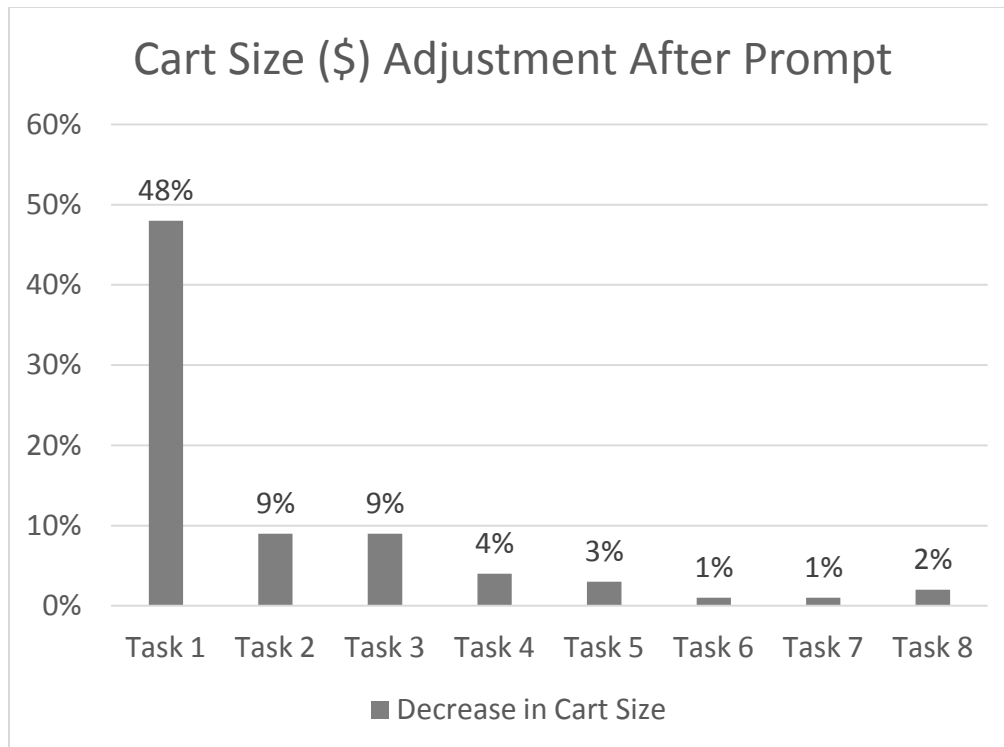
What about overstatement? Don't respondents exaggerate how much they will spend? Yes, at the 2017 ART Forum Hardt-Allenby showed for the same group of people, dollars spent on a conjoint task were about 50% higher than they were in store.

	Conjoint Data (SP)			Transaction Data (RP)		
	Mean	Median	SD	Mean	Median	SD
No. of Purchase Incidents	12.00	12.00	0.00	19.12	15.00	12.20
Units per Incident	3.45	2.92	2.41	1.79	1.65	0.69
Units per Prod.	1.74	1.50	1.01	1.34	1.25	0.33
Max \$ Spent per Incident	20.39	17.62	12.24	13.44	11.99	6.33
Unique Prod. per Incident	2.05	1.69	1.07	1.33	1.20	0.41

Hardt-Allenby (2017 ART Forum)

The disparity in spending didn't seem to have an impact on managerial inference. They showed that you can uncover the same choice process for both conjoint and in-store shopping. Regardless, this is still something to keep an eye on.

In the macaroni and cheese study, we allowed consumers to edit their cart and cart sizes (dollars) were adjusted downward by 48% in the first task. The adjustment tapered off dramatically in subsequent tasks.



Ideally, we'd like to compare the spending amounts with the transactional data, but we did not have that level of granularity in the syndicated sales data.

WHY THE STANDARD MODEL DOESN'T WORK HERE

In the world of choice models, the most common model applied is the multinomial logit model (MNL). This model does an exceptional job under a wide variety of circumstances and is often referred to as the standard model.

The MNL requires the response data to be a single choice (including none). For macaroni and cheese, it is very common for people to buy multiple boxes/cups. Also, they can seek variety across different flavors and brands. Forcing respondents to only choose one option would go against the economic theory of consumer choice that states consumers will choose the option(s) that maximize utility.

Volumetric conjoint loosens the restrictions of a single choice. But carries with it a more complicated model to understand and forecast the complexities of consumer choice.

MODEL OVERVIEW

A new R package is available to do the modeling. The package is called VDMDU. The package author, Nino Hardt, has done an amazing job making the modeling fast and robust (Kim and Hardt 2016).

The model is based on Direct Utility Theory and brings in some new concepts (compared to the standard model) to help understand the consumer choice process. The two new features are the budget constraint and satiation.

Budget Constraint

The model has an explicit parameter for budget. With this parameter, products or alternatives can be thought of as competing for the individual's budget. The forecast spending is constrained such that it will never exceed the budget constraint. Any option that exceeds the budget constraint will have demand of zero.

The estimation of the budget constraint is carefully modeled such that its lowest possible value corresponds to the highest spending value in any of the conjoint choice tasks. This guarantees that the optimal in sample predictions are feasible.

In simulation you'll see that lowering price frees up budget and consumers might buy more units of that product.

Satiation

sā'shē-āt'

1. To satisfy (an appetite, for example) fully
2. To provide (someone) with more than enough

The standard model has a completely reasonable assumption that utility can be maximized with just one option. That reasonable assumption works well in most choice situations. In some categories, like food and beverage, it is unrealistic. In these categories, consumers derive more utility from more units.

The model's satiation parameter governs the diminishing rate of utility for additional units as product utility increases or price decreases.

Other Observations about the Demand Model

There is an upper bound on simulated market revenue. Each respondent's forecasted spending will not exceed their estimated budget constraint. This upper bound does not apply to units or profit—just revenue. This is different from the standard model that has an upper bound of share of choice.

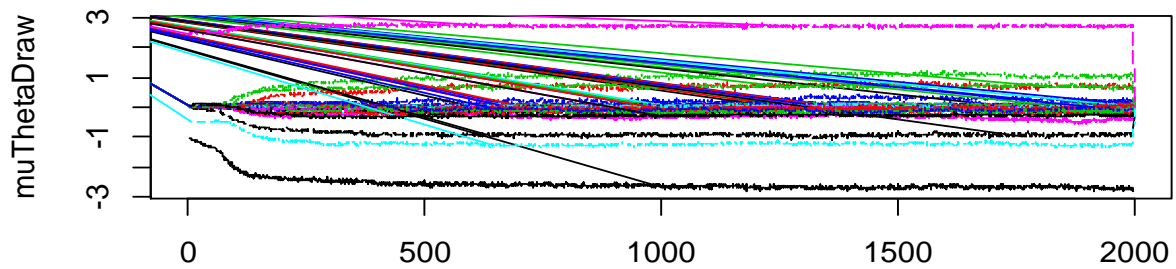
Forecasted quantity demanded increases with utility, but at a diminishing rate. If you make a product better, an individual might choose more of it.

There is no price parameter in the model. The actual product price is part of the forecasting equation. Price sensitivity must be derived in simulation. Unlike the standard model, there is no chance of reversed price sensitivity.

The derivation of the utility and demand functions are beyond the scope of this paper. See the academic work for the formulas and theoretical background. Howell (2013) provides a good overview.

MODEL OUTPUT

This class of models converges very fast compared to the standard model. Below is a traceplot of the posterior means. Every 10th draw is plotted. The lines become horizontally flat (a sign of convergence) relatively quickly.



The model coefficients made sense when compared with the market sales data.

Attribute	Level	Mean	2.50%	97.50%
Brand	Intercept	-2.55	-2.75	-1.35
	Velveeta	0.62	0.01	0.81
	Kraft	0.95	0.01	1.17
	Annie's	0.21	0	0.36
	Cracker Barrel	-0.11	-0.2	0
	Pasta Roni	-0.29	-0.47	-0.02
	Private Label	-0.15	-0.26	0.01
	Pepperidge Farm	-0.16	-0.27	-0.01
	Horizon			
Size	7.25 ounce box	0.64	0.07	0.77
	2 ounce cup			
Type	Elbow Macaroni	0.1	-0.01	0.18
	Shells	0.02	-0.05	0.08
	Spiral	-0.03	-0.11	0.05
	Fun Shapes			
Flavor	Cheddar	-0.07	-0.14	0
	3 Cheese	0.05	-0.02	0.12
	White Cheddar	-0.11	-0.18	0
	Sharp Cheddar	-0.03	-0.09	0.04
	Creamy	0	-0.07	0.08
	Wisconsin Cheddar			
Cheese Form	Powder Cheese Packet	-0.04	-0.09	0.02
	Liquid Cheese Packet			
Preparation	Stove Top	-0.04	-0.1	0.02
	Microwave	0	-0.06	0.06
	Stove Top or Microwave			
Health Claim	No Claim ""	0.04	-0.03	0.12
	Organic	0.03	-0.03	0.1
	Gluten Free	0.02	-0.05	0.09
	Low Calories	-0.04	-0.11	0.03
	10 grams of protein			
Pack Multiple	Single pack	-0.87	-1	-0.11
	4 pack	-0.02	-0.11	0.11
	8 pack	-0.09	-0.2	0.12
	12 pack	-0.05	-0.18	0.15
	18 pack			
	ln(gamma)	-1.18	-1.31	-0.49
	ln(E)	2.7	2.51	2.77
	ln(sigma)	-0.26	-0.34	0.08

For practitioners used to the standard model, it can be surprising to see there is no coefficient for price. The actual/simulated price is in the demand function and we do observe downward sloping demand curves.



Interestingly, for practitioners who are used to seeing occasional positive price coefficients in the standard model, this model will never have price reversals because lower price will always lead to more demand. The relationship is baked into the assumptions of the model.

NOTES ON EXPERIMENTAL DESIGN

The purpose of the experimental design is to efficiently reduce uncertainty. Most often d-efficiency is used. Its objective function is to reduce the uncertainty in the parameter estimates. Other measures of efficiency (V, A, G, etc.) are possible.

For the standard model a random design is essentially equivalent to a d-optimal design that assumes all parameters are zero.

A random design for a volumetric model will struggle to accurately capture the new parameters for budget and satiation. For example, to get a good read on satiation, you will want tasks that push the boundaries on utility and price to best help the satiation parameter know where that fully satiated line is.

You should either construct a model-specific efficient design, which is understandably difficult, or consider adding in some additional tasks to better model the relationship between volume and the model parameters at the extremes of price and utility.

RECOMMENDATIONS AND CONCLUSION

Volumetric models of demand for conjoint analysis are still very young. The model shows a lot of promise for managerial inference when the standard model assumptions don't fit.

The model is a natural fit for the food and beverage categories. It could be appropriate for entertainment categories like movies and theme parks. Any time consumers would regularly pick

multiple options that are competing for the consumer's budget, the direct utility model may be more appropriate than the standard model.

FUTURE RESEARCH

In simulation we saw more price sensitivity than expected. Perhaps too much as the model suggested to optimize revenue, price should be set at the lowest value tested. We noted that the demand curves flattened out a bit when the satiation parameter, γ , was higher. More work is needed to understand the right level of price sensitivity and the role of experimental design on the parameter estimates.



Jake Lee

REFERENCES

- Howell, John R. and Allenby, Greg M., Choice Models with Fixed Costs (March 5, 2013). Fisher College of Business Working Paper.
- Howell, John R.; Lee, Sanghak and Allenby, Greg M., Price Promotions in Choice Models (May 28, 2013).
- Kim, Jaehwan; Allenby, Greg M. and Rossi, Peter E., Volumetric Conjoint Analysis (May 2004).
- Kim, Youngju; Hardt, Nino; Kim, Jaehwan and Allenby, Greg M., Conjunctive Screening in Models of Multiple Discreteness (April 12, 2016).
- Satomura, Takuya; Kim, Jaehwan and Allenby, Greg M., Multiple Constraint Choice Models with Corner and Interior Solutions (August 10, 2010).

A COMPARISON OF VOLUMETRIC MODELS

THOMAS C. EAGLE

EAGLE ANALYTICS OF CALIFORNIA, INC.

JORDAN LOUVIERE

UNIVERSITY OF SOUTH AUSTRALIA

TOWHIDUL ISLAM

UNIVERSITY OF GUELPH, CANADA

INTRODUCTION¹

Volumetric models have a varied history in marketing. Some practitioners and academics avoid them. Volumetric models attempt to predict the number of units of a product or alternative a consumer would buy. Some argue the respondents are incapable of giving accurate volumetric responses to experimentally designed tasks (Huber, 2018 discussion of Jake Lee's presentation). Others say the data are too inconsistent and too crazy to model in any reasonable fashion. Some even refer to bad experiences by previous researchers to model and predict volume (e.g., Turbo Choice Modeling discussions). Nevertheless, building volumetric models from data collected in an experimentally designed surveys has become the topic of recent papers at the Sawtooth Conference and ART Forums over the years (e.g., Eagle, 2010; Garratt and Eagle, 2010; Howell and Allenby, 2012). Recent academic papers include the hierarchical Bayesian volumetric models proposed by Kim, Allenby and Rossi (2007) and, more recently, that of Pachali et al., (2017) and Hardt et al. (2017).

The motivation of this presentation is two-fold:

1. To promote the use of volumetric models which are now easier to estimate than ever before.
2. To examine the patterns of substitution inherent in three different approaches to modeling volumetric data.

THE MODELS COMPARED

Joint Discrete/Continuous Model

Eagle in (2010) proposed a series of models that consist of two components: a share model explicitly designed to capture the patterns of substitution among alternatives in a given scenario; and a volumetric model that uses the predictions from the share model to predict either total task volume or alternative specific volumes.

By substitution we are referring to the change in the alternatives' share (and units) of volume that occurs when the attributes of the alternatives change. For example, if we raise the price of product A, we would expect the volume of product A to drop. If that loss in volume goes to another product(s), then that change in the other products' volume is called substitution. The multinomial logit model is an explicit choice model that directly predicts substitution. In the case of the

¹The design of the data used in the study was done by Louviere and Islam. Louviere provided some input to the modeling and Islam commented on the material presented. Otherwise most of the work, including the writing was done by Eagle.

multinomial logit model, the substitution that is predicted is a constant relative change in the predicted probabilities of the non-changing alternatives.

Originally estimated sequentially (i.e., the share model, then the volumetric model), these joint models can now be estimated jointly using hierarchical Bayesian methods.

The first component of the model may consist of *any* valid specification of a probabilistic choice model. Typically, this is an MNL model, but any form of a choice model can be used. This component of the model is explicitly designed to capture the patterns of substitution among the alternatives in a scenario. The volumetric data is converted into a share of volume and those shares used to fit the share model component.

The volumetric component uses the predictions from the share component as inputs into the volume model. In the case of the total task volume model, this input is the natural log of the net expected utility of the set of alternatives (e.g., $\text{Ln}[\text{Denominator of MNL model}]$). In the case of the alternative-specific volume model, the inputs are the predicted probabilities of the alternatives. Additional terms can be added into the volumetric model (such as bias adjustments: Train, 1986) and a common set of upper level model covariates may be used.

Estimated simultaneously using hierarchical Bayesian methods (specifically RSGHB), the parameters of these two components now share a common parameter covariance matrix and the error components can be allowed to be correlated. Typically, one uses the lower-level posterior draws to predict (or point estimates of such a model), but one can also predict using the upper-level model parameters. See Eagle (2010) for details on the model formulas.

This is a descriptive type of volumetric model, pure and simple. It is engineered to predict volume data without any direct ties to specific theories of demand. It is based upon the concept of indirect utility rather than direct utility. However, there is previous academic work using such models (Hanemann, 1984; Train, 1986; and Hausman, et al., 1995), but not in the context of fitting hierarchical Bayesian models.

The Hardt-Allenby Model

Hardt and Allenby (2017) presented an elegant direct utility form of a volumetric model which is an extension of the earlier model presented by Howell and Allenby in (2012). Mathematically derived and quite elegant, the authors have built a model that is comprised of three components: the consideration of an alternative at a given price; a budget constraint; and a satiation component. The consideration component addresses whether the respondent would even consider an alternative at it given price. This naturally affects the set of viable, or considered, alternatives in any scenario. The budget constraint is part of the consideration component, but it also affects whether the respondent would buy anything at all. This is captured through the use what is called the outside good (e.g., the None alternative in choice modeling parlance). All substitution among alternatives is captured through the outside good. There is no direct, alternative-to-alternative, modeling of substitution. The satiation parameter enables the model to capture the diminishing benefit of continually increasing volume. That is, there is satiation in demand. See Hardt, et al. (2017) for details of the model formulas.

Nino Hardt has made the estimation of this hierarchical Bayesian model very easy. He has developed R code that is very fast and easy to setup, to fit, and predict using the model. One can acquire the R code direct from Hardt. The model includes covariates in the upper level and

estimates individual-level parameters. Predictions can be generated using the upper-level model or using lower-level posterior draws.

The Hardt-Allenby model is an explanatory model of individual-level demand (Allenby discussion of this presentation, 2018). As such, it is designed to explain the behavior we see in addition to predicting it. It is not, however, optimized for prediction. That is not its main purpose. Nevertheless, it does predict quite well in the examples shown at conferences and in papers.

Our focus was on the ability of the model to capture patterns of substitution among similar alternatives. We are skeptical of the ability of the model whose sole source of all substitution is through the outside good and budget constraint. To further elaborate, the main component of the Hardt-Allenby model is a direct utility function that is purely a function of the alternative's attributes and the alternative's price. If that price exceeds the budget constraint (a parameter in the model), then volume drops. That volume goes to the outside good, which includes not buying anything at all and all other alternatives. We want to see if the model works from the managerial perspective.

The Latent Class Poisson Model with Cross Effects

The final model tested is that of a Poisson model. Poisson, or count, models are often used to model volume. They are designed to model data that ranges from 0 to a maximum volume. The models are easily fitted in a variety of ways including regression-like aggregate models, latent class models, and hierarchical Bayesian models. We chose to use a latent class formulation for a change of pace. These models have as a dependent variable, the number of units assigned to an alternative. The independent variables are the attributes of the alternative under consideration. As such there is nothing in the model to capture the substitution of one alternative on the other.

To remedy this lack of substitution in the classic Poisson model we add cross effects from the other alternatives into the specification of each product's volumetric model. Rather than use every attribute as a cross effect, we limited the specification to only the two attributes that have the biggest impact on volume. Using too many cross-effect attributes leads to severe overfitting issues in these models.

One important issue with any cross-effect model is what happens when you delete or add alternatives to a task. Deleting an alternative is simple enough: one simply drops the cross effects associated with the deleted alternative in the remaining alternatives' volume models. Adding an alternative presents a problem. Does one add more cross-effect terms to the existing alternatives' model specifications? If so, what should the new parameter values be? Or, should one replicate the existing alternatives' models using the new alternative's attribute values in the cross-effect terms of the replicated models, and average across the existing alternatives' volume predictions? This issue is not discussed in the literature nor in Sawtooth Software's previous conferences. In this paper we tried both methods and the results were approximately the same. We ended up using replicated cross-effect terms with their already estimated parameters to capture the effect of adding new alternatives to a task.

THE DATA

The data is from a study being conducted by the authors examining volumetric modeling across four different readily consumed products. The specific data used in this paper is for canned tuna.

Respondents are part of an IRI panel recording their actual purchase of the four products over a specific time horizon consisting of three waves:

1. Wave 1. The respondents were given a stated preference, volumetric, set of tasks in addition to other survey information,
2. Wave 2. The respondents' purchases of the same product were monitored after conducting the volumetric task.
3. Wave 3. A follow-up to wave 2 in which some respondents responded to a new set of volumetric tasks after their real purchases were monitored.

There are 738 respondents with complete data across all waves for the canned tuna dataset. The canned tuna task is shown in Figure 1. The task is part of a designed volumetric stated preference task in which we asked for the number of units of canned tuna each respondent in a closed form response (0, 1, 2, . . . , to more than 6 units). Respondents could assign units to multiple alternatives in the task. If they selected zero for all alternatives in the task, then their total volume was zero. The context for the task was the respondent's next regularly scheduled shopping trip.

Figure 1. The example stated preference, volumetric task given to all respondents in waves 1 and 3.

Task Instructions					
<p>In this section of the survey we will show you 12 different supermarket shelf displays for brands of canned tuna. Each shelf displays several brands, and each brand is described by several product features. The purpose of this section of the survey is to better understand how consumers like you evaluate and purchase canned tuna. There are no right or wrong answers. The only thing that matters is that you try your best to tell us what you would actually buy if the brands described were available the next time you shop for canned tuna.</p> <p>All you need to do is evaluate the brands offered in each shelf display and tell us how many of each brand you would be likely to buy the next time you go to your local retail outlet for canned tuna. All you have to do is click on the quantity below each brand to tell us how many cans you want to buy.</p> <p>An example of how to answer the questions is shown below</p>					
	StarKist	Bumble Bee	Chicken of the Sea	Store Brand	Any Other Brand
Tuna Type	Tuna	Albacore	Albacore	Tuna	Tuna
Packed in	Water	Oil	Water	Oil	Oil
Form	Solid	Solid	Solid	Chunk	Solid
Size	12 oz	6 oz	12 oz	12 oz	6 oz
Price / oz	\$0.30	\$0.25	\$0.25	\$0.21	\$0.51
Price per can	\$3.66	\$1.49	\$3.04	\$2.47	\$3.07
How many cans of each would you buy (Check ONE BOX in each COLUMN to the right)?	Select ONE ANSWER per brand below.				
	<input checked="" type="radio"/> 0	<input type="radio"/> 0	<input type="radio"/> 0	<input type="radio"/> 0	<input checked="" type="radio"/> 0
	<input type="radio"/> 1	<input checked="" type="radio"/> 1	<input type="radio"/> 1	<input type="radio"/> 1	<input type="radio"/> 1
	<input type="radio"/> 2	<input type="radio"/> 2	<input checked="" type="radio"/> 2	<input type="radio"/> 2	<input type="radio"/> 2
	<input type="radio"/> 3	<input type="radio"/> 3	<input type="radio"/> 3	<input checked="" type="radio"/> 3	<input type="radio"/> 3
	<input type="radio"/> 4	<input type="radio"/> 4	<input type="radio"/> 4	<input type="radio"/> 4	<input type="radio"/> 4
	<input type="radio"/> 5	<input type="radio"/> 5	<input type="radio"/> 5	<input type="radio"/> 5	<input type="radio"/> 5
	<input type="radio"/> 6	<input type="radio"/> 6	<input type="radio"/> 6	<input type="radio"/> 6	<input type="radio"/> 6
	<input type="radio"/> More than 6	<input type="radio"/> More than 6	<input type="radio"/> More than 6	<input type="radio"/> More than 6	<input type="radio"/> More than 6

In the above example Store Shelf Display the person has selected to buy 1 can of Bumble Bee 6 oz Albacore solid tuna in oil, 2 cans of Chicken of the Sea 12 oz Albacore solid tuna in water, 3 cans of Store Brand 12 oz Tuna chunk in oil and 0 for all the other options.

Please click ">>" to proceed to the first of 12 shelf displays and tell us how many (if any) of each product you would buy.

The tasks were generated from a 5 (brands) x 2^4 x 4 design. The attributes were:

- Brand: 5 levels—StarKist, Chicken of the Sea, BumbleBee, A store brand, and Any other brand
- Type of tuna: 2 levels—Albacore or Regular
- Packed in: 2 levels—Oil or Water
- Form: 2 levels—Chunk or Solid
- Size: 2 levels—6 oz or 12 oz can
- Price: 4 levels—unique levels assigned to each brand derived from using IRI data collected prior to wave 1

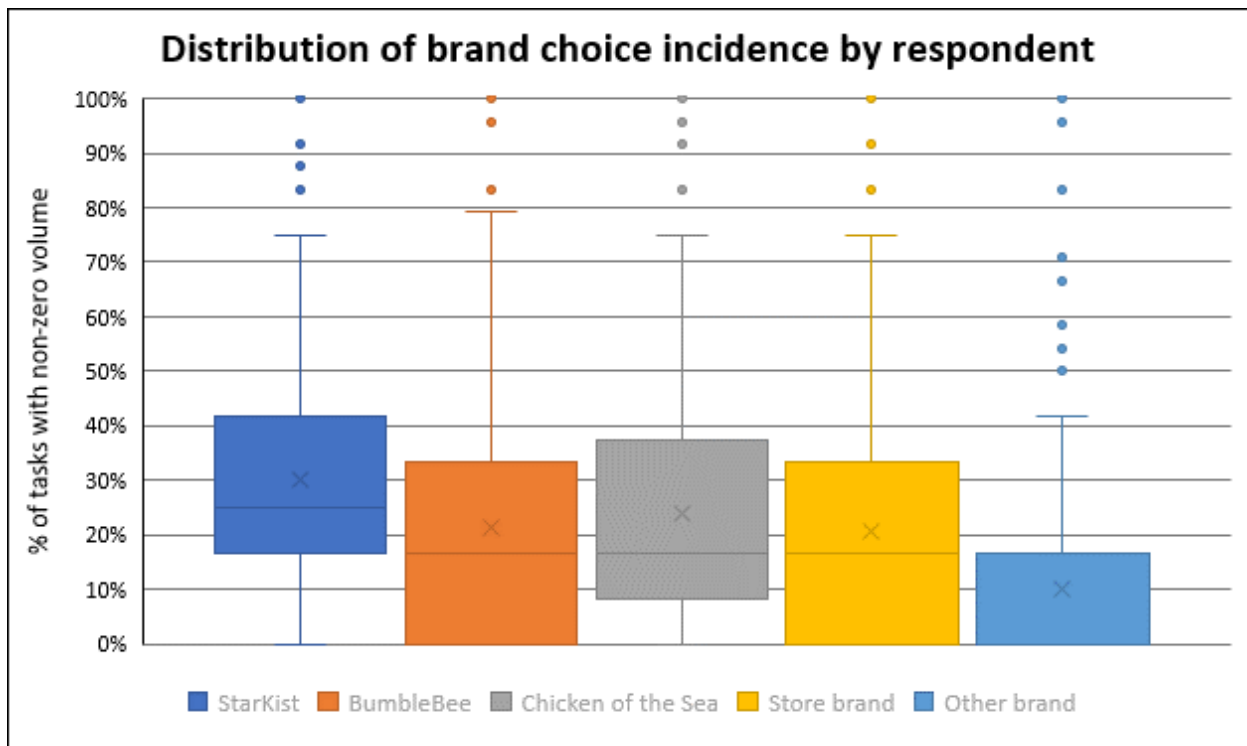
The final design consisted of alternative-specific attribute levels constructed using a fractional factorial design. The design consisted of 8 blocks of 8 tasks each. Each respondent was randomly assigned to one of the 8 blocks. Tasks were randomly assigned to each respondent within each block. All respondents completed the same 4 holdout tasks.

In this market, the expectation is that the products are substitutable and removing one brand, adding an additional SKU, and changing price while holding all else constant should lead to not only some change in the volume of the other alternatives/brands, but also a change in total volume purchased. There is likely some latent demand given the tasks show only 5 SKUs at any given time. So, adding a new SKU would likely increase total volume, but it would likely take some volume away from the original set of brands. Removing a brand should lead to the converse effect.

An important point about these data and the design is that the number of alternatives remained constant across all tasks. As such, we have no data to explicitly compare to the model predictions when we add or remove products. Any substitution we predict while adding or removing an alternative is purely a result of the model estimated and its parameters. This will become an important point in the comparisons later.

Figure 2 below shows the distribution of brand volume incidence by respondent (i.e., every time a respondent assigned a non-zero volume to a brand they are given a 1; else a zero) across the 8 estimation tasks they saw. There is a good distribution of brands chosen across the tasks respondents saw. Only 6% of the respondents chose the same brand 100% of the time across their 8 tasks. 62% of the respondents chose 4 or more brands across all their tasks.

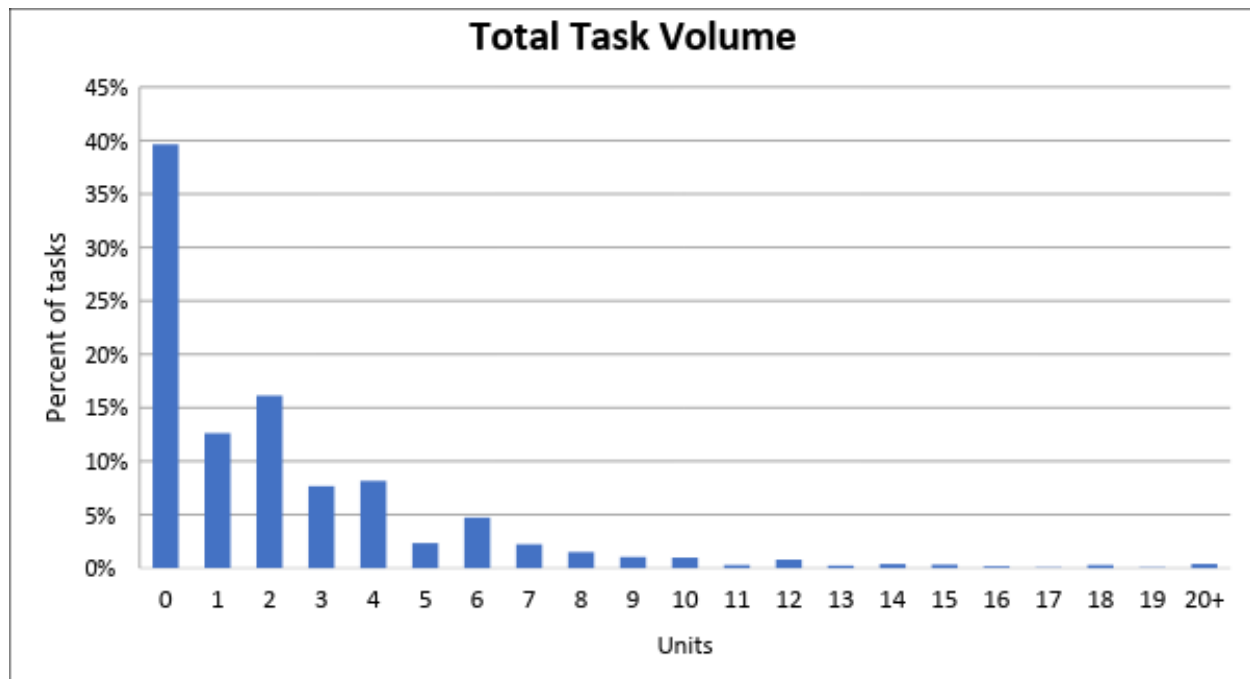
Figure 2. Distribution of Brand Choice Incidence by Respondent



The volumes assigned across tasks also showed a good distribution. 39.6% of the tasks had a zero total assigned volume. 34.9% of the tasks had all volume assigned to a single alternative in the task. 25.5% of the tasks had volume assigned to 2+ alternatives.

Figure 3 depicts the distribution of total task volumes assigned across tasks seen by the respondents. It shows a clear spike at zero and minor spikes at 2 and 6 units. It resembles a Poisson distribution.

Figure 3. Distribution of Total Tasks Volume Across All Tasks



ESTIMATION

The joint/discrete model and the Hardt-Allenby model were both estimated using hierarchical Bayesian methods. The R package RSGHB was modified to handle the simultaneous estimation of the joint discrete/continuous volumetric model by rewriting the likelihood function. The Hardt-Allenby model was estimated using the highly refined R routines built by professor Nino Hardt. The latent class Poisson model was estimated using the Latent Gold package.

All models were built using generic parameters across alternatives. That is, the impact of type of tuna is the same across all brands. Price was also estimated as a generic effect but transformed using the natural logarithm. The design allowed alternative-specific parameters to be estimated, but, for simplicity's sake, we used generic parameters.

Both the joint discrete/continuous and Hardt-Allenby models were estimated using 100,000 burn-in iterations. 10,000 post burn-in iterations, saving every 10th iteration, were used for the model predictions in the subsequent model comparisons. We used the saved 1,000 posterior draws for all comparisons. We did predict volumes using the saved individual-level posterior parameters and using draws from the upper-level model in order to compare them. The predicted values for volume across both prediction methods were indistinguishable from one another. This is likely because we did not include any covariates in the upper level of the model estimation.

The latent class Poisson model estimated using Latent Gold was conducted using all 5 brands' volume models. Each brand had its own set of unique constants, generic main effect parameters, and alternative-specific cross effects of every other brand in their model. This enabled the latent class program to build a single set of segments across all brand volume models. The solution showed a continuously improving BIC criteria up to 20 segments—where we stopped estimation. This suggests an extreme amount of heterogeneity in the data.

As most practitioners and clients have difficulty with more than 6-8 segments, we examined how the BIC criteria shrank as the number of segments grew. A scree plot of the BIC criteria across the number of segments showed a bend at six segments. Six segments were ultimately used for the latent class Poisson model. The model parameters were consistent with respect to the main effects with those of the HB estimated models. However, the cross-effect parameters were a mix of positive and negative signs.

In the case of every model, all parameters were included in the final set of models. Dropping those terms that might be considered insignificant based upon the conventional criteria for both the HB and latent class was not done—as would normally be the case in model prediction.

PREDICTION TO ESTIMATION DATA AND HOLDOUT TASKS

Estimation Data Fit Measures

Tables 1 and 2 show some key prediction metrics from the models. In the tables and figures below the following labels are used:

- JDC for the joint discrete/continuous volumetric model
- Hardt-Allenby for the Hardt-Allenby direct utility volumetric model
- LC Poisson for the latent class cross-effect Poisson model

Table 1. R-square Values of the Actual Volumes vs. the Predicted Volumes

Total Volume		Volume by Alt	
JDC	0.61	JDC	0.47
Hardt-Allenby	0.53	Hardt-Allenby	0.48
LC Poisson	0.26	LC Poisson	0.29

Examining the R-square values across models on the estimation data (Table 1) shows that the JDC model performs better in predicting the total volume across all alternatives across tasks. When examining the alternative specific volumes predicted across tasks the JDC model and the Hardt-Allenby model predict about the same. The latent class Poisson model is the clear runner-up, but it has many fewer parameters than either HB model.

Table 2. Mean Absolute Error in Units of Volume

Total Volume		Volume by Alt	
JDC	1.417	JDC	0.459
Hardt-Allenby	1.566	Hardt-Allenby	0.475
LC Poisson	2.287	LC Poisson	0.571

Table 2 shows the mean square error of estimation for each model for the total and alternative-specific volumes. The JDC model predicts the actual volume data in the estimation data set for both

total task and alternative-specific volumes better than the other models. The difference between the JDC and Hardt-Allenby model is small, whereas the latent class Poisson model again trails in terms of internal model fit.

Holdout Task Prediction

Tables 3-6 show the predicted total volumes for each model to the 4 holdout tasks. Each table shows the actual vs. predicted total task volumes, the percentage difference in the values, and the mean absolute error across the 5 alternatives in each task. The general conclusion is that the Hardt-Allenby model performs better in both the estimation of total task volume and alternative specific volumes across all 4 holdout tasks.

Table 3. Holdout Predictions to Holdout Task 9

Holdout 9	JDC	Hardt-Allenby	LC Poisson
Predicted Total Volume (Actual: 1047 units)	1082	996	782
% Diff	+3.3%	-4.9%	-25.3%
MAE (units across 5 alts)	47.97	30.93	53.03

Table 4. Holdout Predictions to Holdout Task 10

Holdout 10	JDC	Hardt-Allenby	LC Poisson
Predicted Total Volume (Actual: 805 units)	944	693	594
% Diff	+17.3%	-13.9%	-26.2%
MAE (units across 5 alts)	48.92	26.75	43.68

Table 5. Holdout Predictions to Holdout Task 11

Holdout 11	JDC	Hardt-Allenby	LC Poisson
Predicted Total Volume (Actual: 2876 units)	1932	2806	1924
% Diff	-32.8%	-2.4%	-33.1%
MAE (units across 5 alts)	197.76	90.23	190.47

Table 6. Holdout Predictions to Holdout Task 12

Holdout 12	JDC	Hardt-Allenby	LC Poisson
Predicted Total Volume (Actual: 2283 units)	1551	1996	1427
% Diff	-32.1%	-12.6%	-37.5%
MAE (units across 5 alts)	163.67	95.71	171.12

Figures 4 and 5 depict the actual vs. predicted alternative-specific volume and total task volumes for two of the 4 holdout tasks (9 and 11). The Hardt-Allenby model predictions are much better on a consistent basis than either the JDC or LC Poisson models. An inconsistency of predicting total task volume is apparent in the JDC model if one compares the predicted total task volumes for the JDC model across Figures 4 and 5.

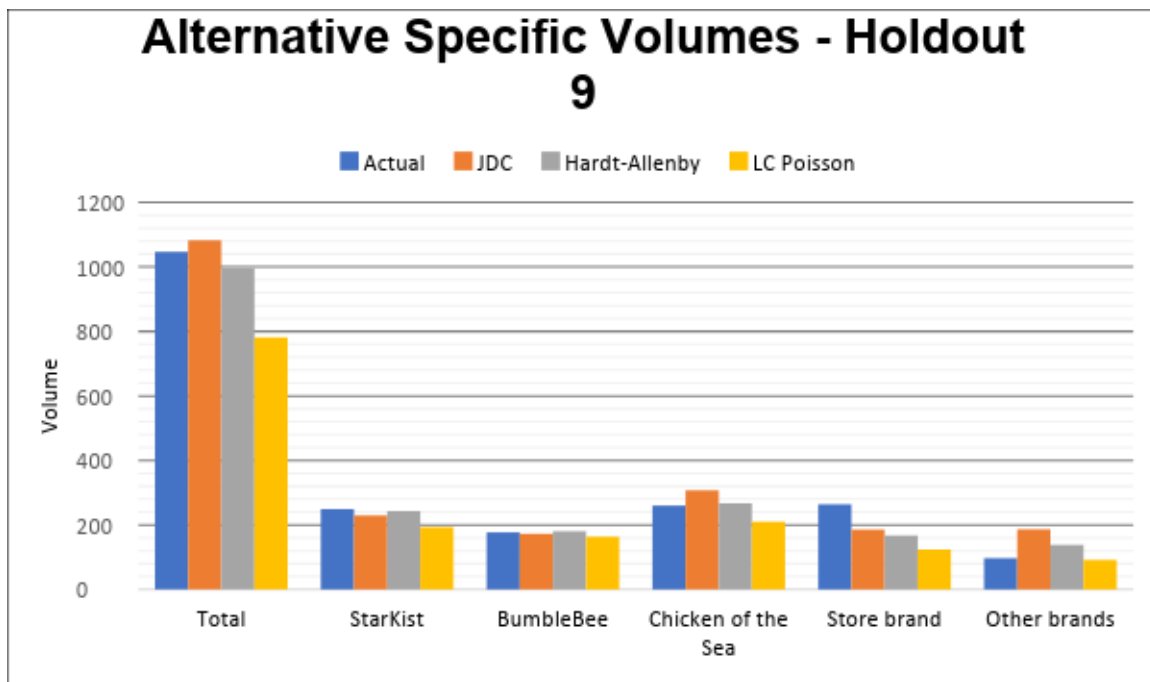
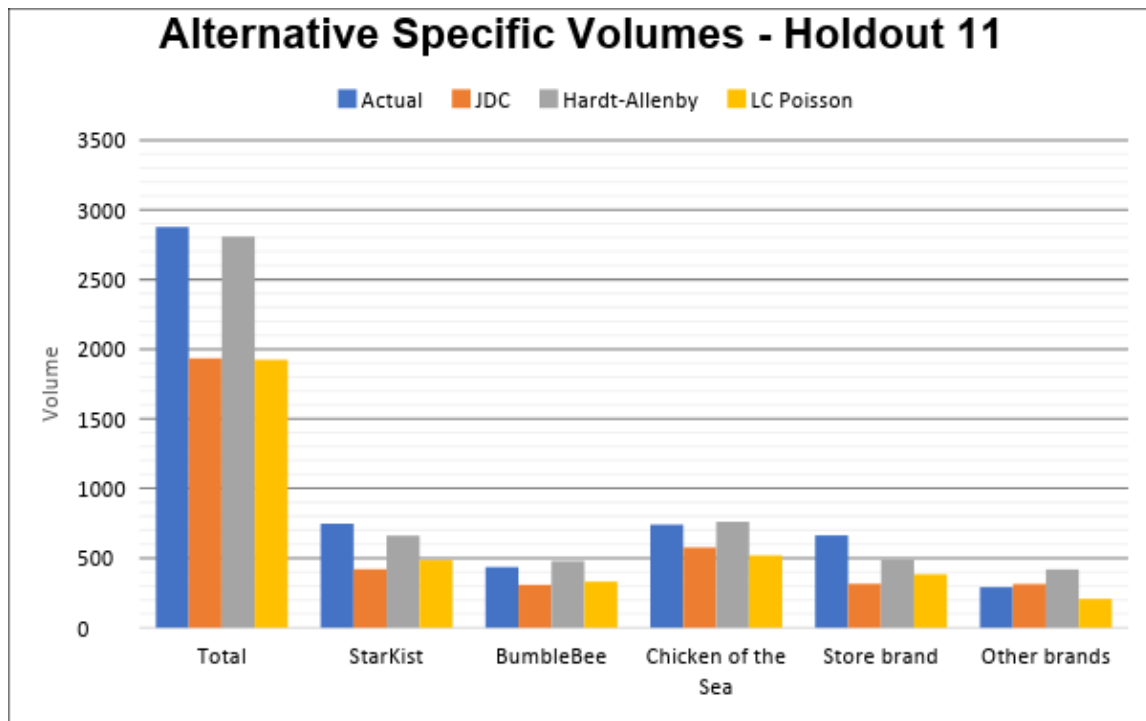
Figure 4. Alternative-Specific and Total Task Volumes for Holdout Task 9

Figure 5. Alternative-Specific and Total Task Volumes for Holdout Task 11



The JDC model does well at predicting the total task volume and alternative-specific volumes for holdout task 9. In fact, all the models do about the same in this holdout task. However, in holdout task 11 the Hardt-Allenby model predicts the total volume and the alternative specific volumes better than the JDC or LC Poisson models.

MODEL COMPARISONS

Four different scenarios are used for model comparison. The emphasis is not about prediction accuracy, rather it is about the face validity of inferences management would make about the consumer behaviors inherent in the model predictions. Because of this emphasis, measures of predictive accuracy are not reported. We simply show charts that would typically be used to present the results of these different market scenarios and report on the differences we see across the model predictions. The four comparisons include:

1. Remove a single alternative from a market (task)—to model a temporary out of stock situation.
2. Add an additional SKU (alternative) for a brand into a market—to model a line extension.
3. Examine the price sensitivity of a brand in a market—examine the own and cross-price sensitivities to a single brand changing its prices.
4. Add an additional SKU to each brand in a market.

Remove a Single Alternative

In this scenario we start with the holdout task number 9. This scenario is outside the domain of tasks shown to respondents. That is, we never showed respondents a task with a single alternative removed; nor did we ever show a task where we added an alternative. All tasks shown to

respondents had the same number of alternatives. Therefore, the predictions are based upon the assumed nature of substitution inherent in the models. Additionally, we do not have raw data to support or refute the predictions of any of the models.

Table 7 shows the attributes and levels for this holdout task. The last two columns show what proportion of the sample gave a non-zero volume to the alternative and the mean volume for those respondents assigning a non-zero volume to the alternative. Every alternative has the same non-price attribute values. The prices are the lowest 12 oz can price for each brand. In this scenario we will drop the Chicken of the Sea 12 oz can from the market.

Table 7. Holdout Task 9

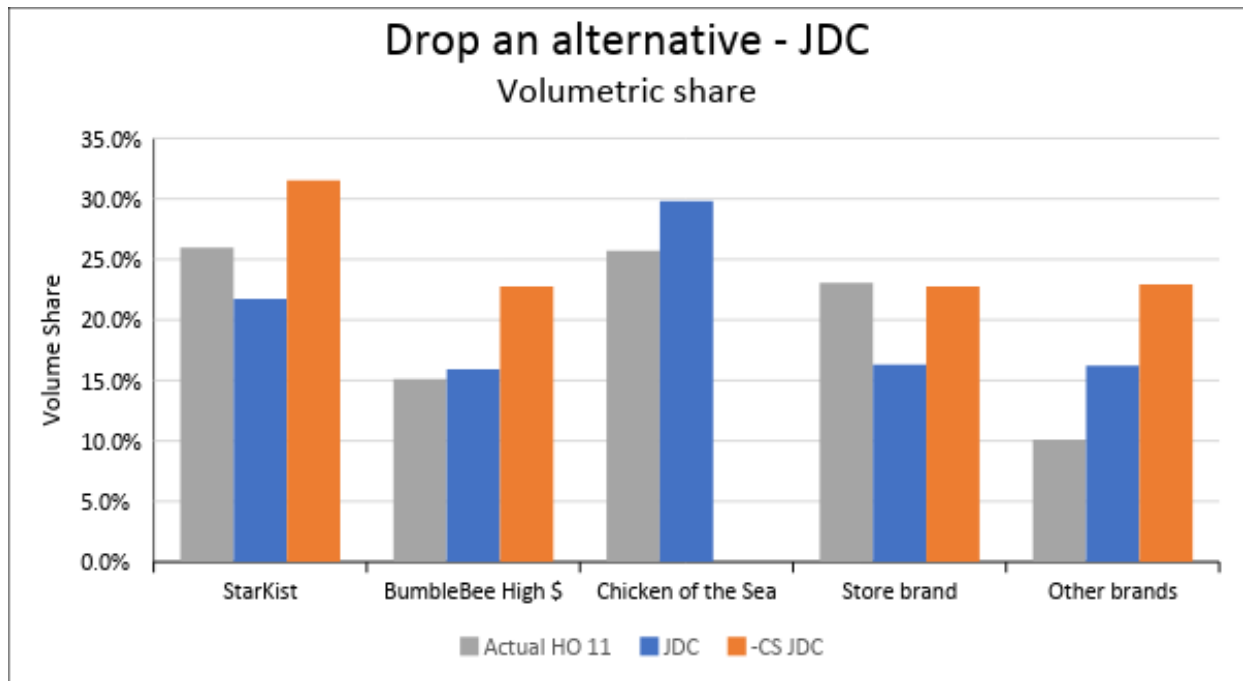
	HO_9						
Brand	Type	Packed-In	Form	Size	Price	% of R choice	Mean volume given choice
StarKist	Albacore	Oil	Solid	12 oz	\$2.99	12.8%	1.80
BumbleBee	Albacore	Oil	Solid	12 oz	\$3.68	8.1%	2.03
Chicken of the Sea	Albacore	Oil	Solid	12 oz	\$2.89	12.7%	1.90
Store brand	Albacore	Oil	Solid	12 oz	\$2.47	11.0%	2.22
Other brands	Albacore	Oil	Solid	12 oz	\$2.80	4.8%	1.87

Figures 6 through 8 present the volumetric share predictions for the joint discrete/continuous (JDC) volumetric model, the Hardt-Allenby direct utility volumetric model, and the latent class Poisson (LC Poisson) volumetric models respectively. In each figure is shown:

1. The predicted volumetric share when all 5 alternatives are in the task (e.g., JDC label), and
2. The predicted volumetric share when the Chicken of the Sea brand is dropped (e.g., -CS JDC label)

The figures look remarkably similar. In each figure the StarKist and Chicken of the Sea brands have the highest actual (1st of the 3 bars) and predicted volumetric shares (2nd of the 3 bars) before Chicken of the Sea is removed. In each figure the removal of the Chicken of the Sea brand results in an increased share of volume for every remaining brand (the 3rd bar in each figure). It appears the largest proportion of the Chicken of the Sea volume goes to StarKist, BumbleBee, and Store brands. In fact, these changes in volumetric share are roughly equivalent to what one would expect if the IIA property held in the models.

**Figure 6. Volumetric Shares for Dropping a Single Alternative
Joint Discrete/Continuous Model**



**Figure 7. Volumetric Shares for Dropping a Single Alternative—
Hardt-Allenby Model**

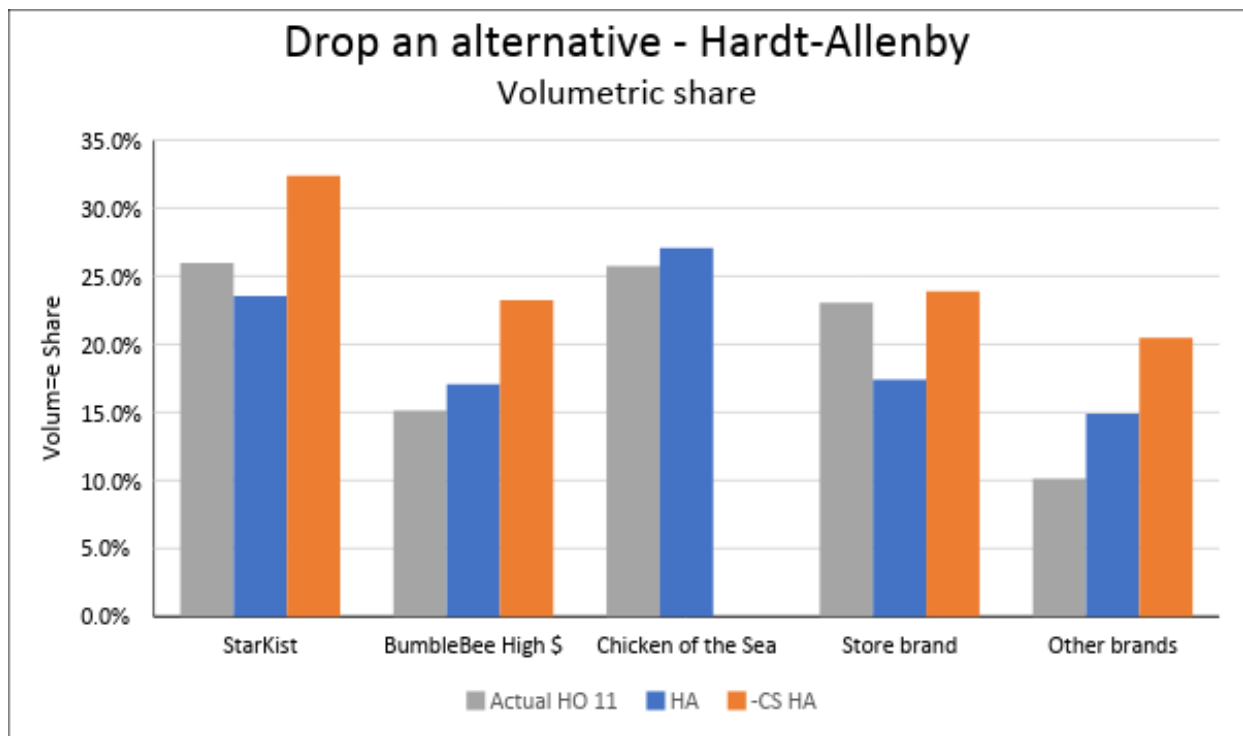
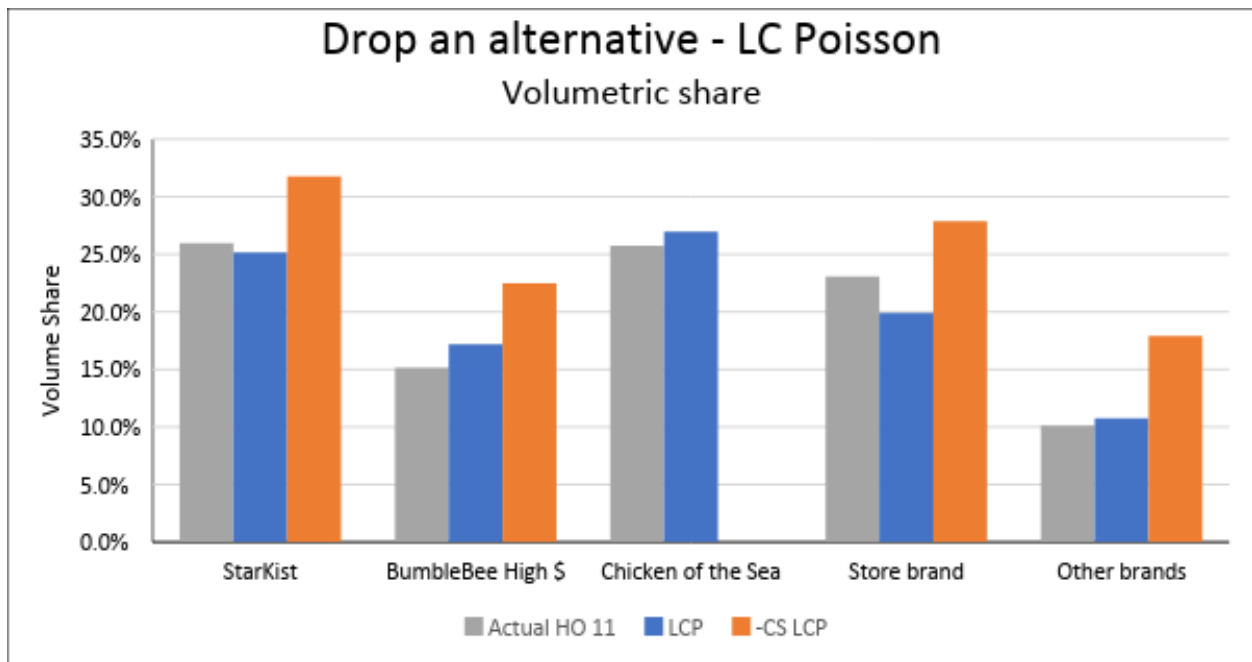


Figure 8. Volumetric Shares for Dropping a Single Alternative—LC Poisson Model



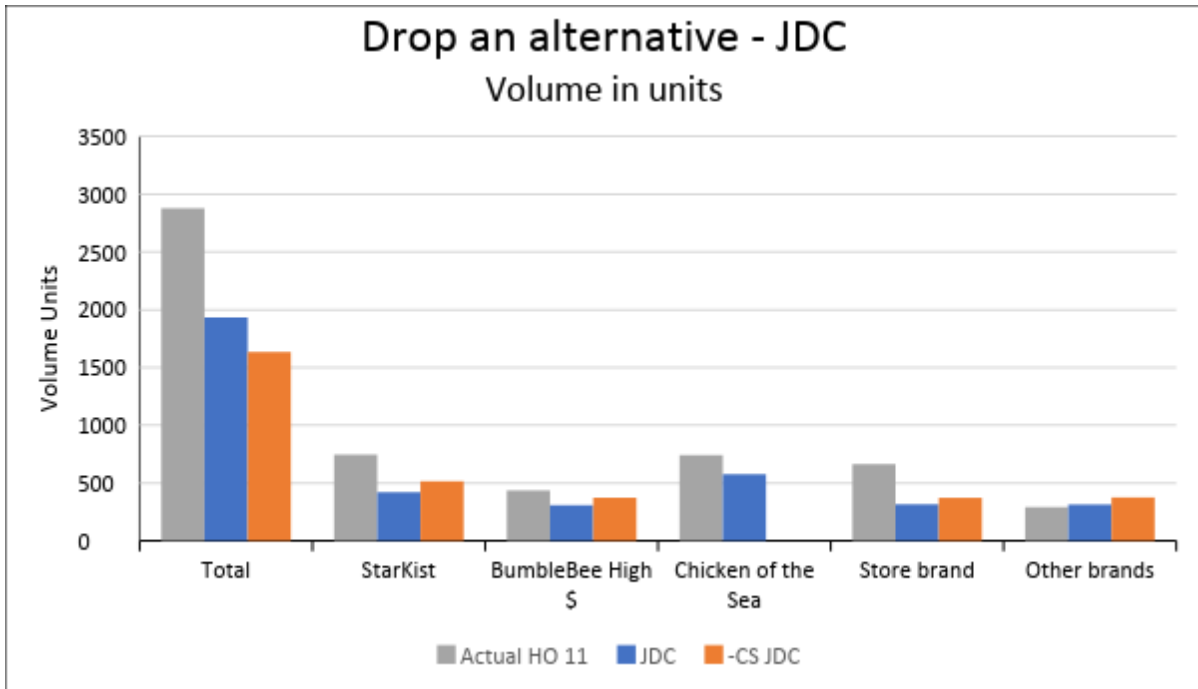
The figures are hiding a key aspect of differentiation across the models. While volumetric share is important to know, many clients wish to know the amount of volume change in units so that revenues and potential profits may be measured. To understand this, we must convert the figures into the number of units predicted before and after the dropping of the Chicken of the Sea brand.

Figures 9-11 depict the number of units predicted for the same three models for the same scenario. Examination across these 3 figures, while expressing the same model results, but in units of volume, reveals some key differences. The first set of three bars depicts what happens to the total task volume. The first bar is the actual task volume. The 2nd bar is the model's predicted total task volume before dropping the alternative. The third bar is the predicted total volume after dropping Chicken of the Sea. The remaining sets of bars are the actual predicted-before and predicted-after volumes for the alternatives in the task.

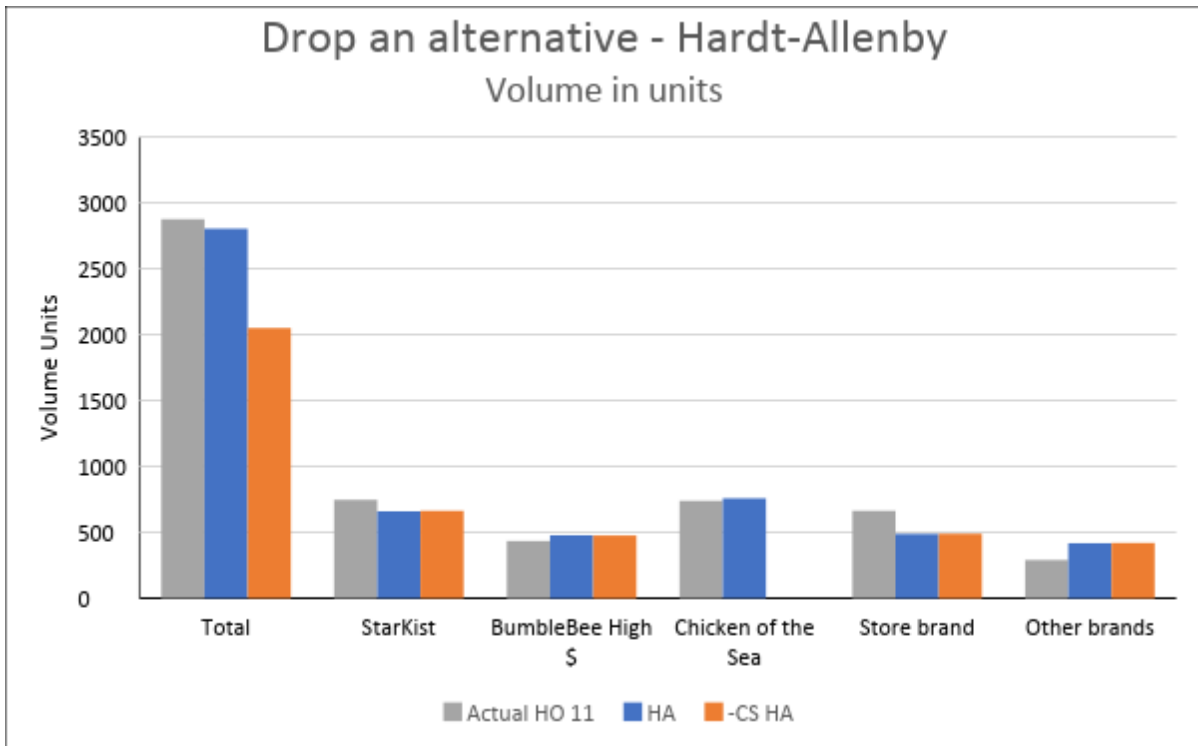
The JDC model (Figure 9) shows some substitution when the Chicken of the Sea brand is dropped. The bars for the remaining alternatives show an increase in volume units, but the net effect is for predicted total volume to drop. The increase in volume to the remaining alternatives is nearly what one would expect given the MNL component of the model which assumes IIA. While it is not exactly IIA, it is very close.

Figure 11, the LC Poisson model, shows a very similar pattern of growth in the volume in units for the remaining brands and a drop in the total volume, but the substitutional pattern deviates from IIA more than the JDC model. StarKist's volume grows by less than 1%; BumbleBee's by ~3%; the store brand by ~11%; and the Other brands by 32%. If IIA held, these percentage changes would be nearly the same.

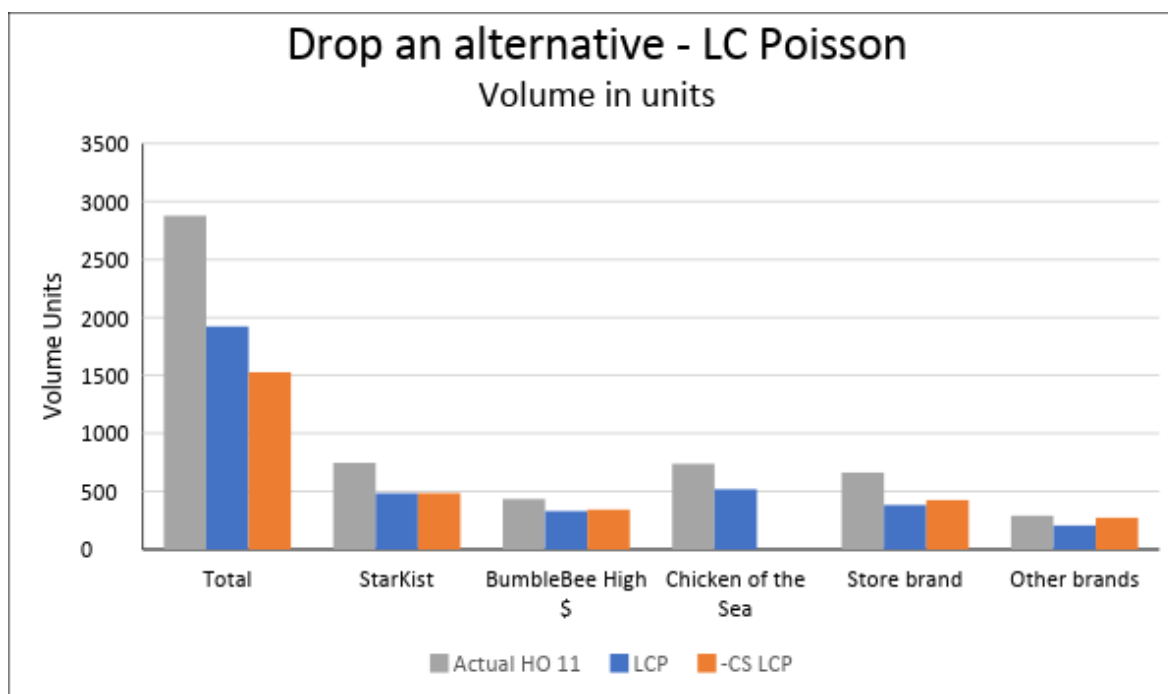
**Figure 9. Volume in Units for Dropping a Single Alternative—
Joint Discrete/Continuous Model**



**Figure 10. Volume in Units for Dropping a Single Alternative—
Hardt-Allenby Model**



**Figure 11. Volume in Units for Dropping a Single Alternative—
LC Poisson Model**



The Hardt-Allenby model (Figure 10) makes substantially different predictions than the other two models. The first set of 3 bars show a drop in predicted total task volume when Chicken of the Sea is dropped. This is consistent with the other 2 models except the amount of total task volume loss is greater between the before and after scenarios (i.e., the drop between the 2nd and 3rd bars of the first set of 3 bars is much greater for the Hardt-Allenby model than the other two models). More remarkable is what happens to the volumes of the remaining alternatives. When Chicken of the Sea is dropped the Hardt-Allenby model predicts that the volumes of the remaining alternatives do not change. The alternatives' volumes stay the same, except for the zeroing out of the Chicken of the Sea brand. The predictions before and after dropping an alternative were within 1-4 units when aggregated across all respondents!

This suggests zero substitution among the alternatives. Dropping Chicken of the Sea results in a drop in total volume, but the volume of all other brands remains constant. All volume is lost. Most managers would have a hard time believing such a story if their expectations assumed some degree of product substitution. The IRI data suggests there is some substitution among the brands, but the Hardt-Allenby model does not support this expectation. Are these results a function of the stated volume data, the design, the context of the choice, or the Hardt-Allenby model itself? That is, is the Hardt-Allenby's model assumption of all substitution, captured by the comparison of each alternative's price to a budget constraint and volume change, going and coming through the outside good, the reason for these results?

Adding a Single Alternative

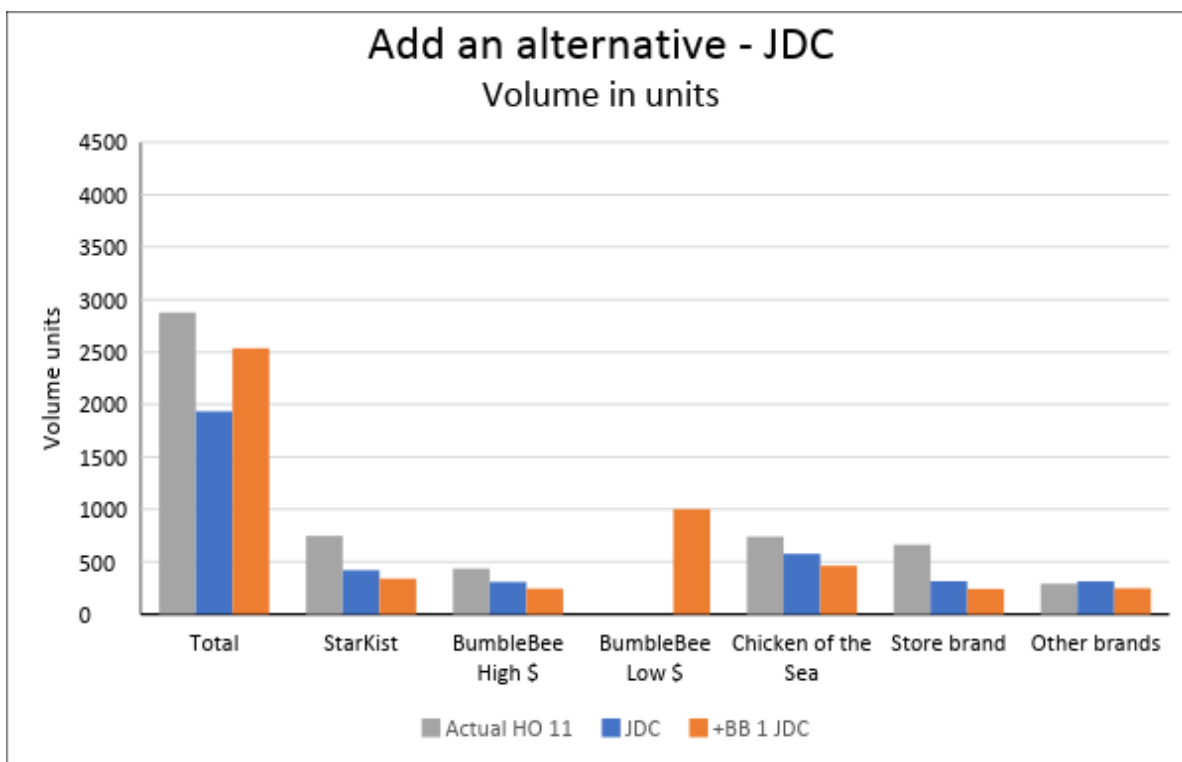
As a counter to removing a single alternative, the second comparison examines what is predicted by the models when one additional SKU is added to a brand. The expectation is the

opposite of removing an alternative: the total task volume should increase and the volume going to the original alternatives should decrease (or stay about the same).

Figures 12 through 14 show the predicted volume in units for the three models. In this case we used holdout task 11 which has all brands with regular tuna type, packed-in water, chunk style, 12 oz cans, and prices at the lowest level of the 12 oz can for each specific brand. We add a single BumbleBee 6 oz can at its lowest price.

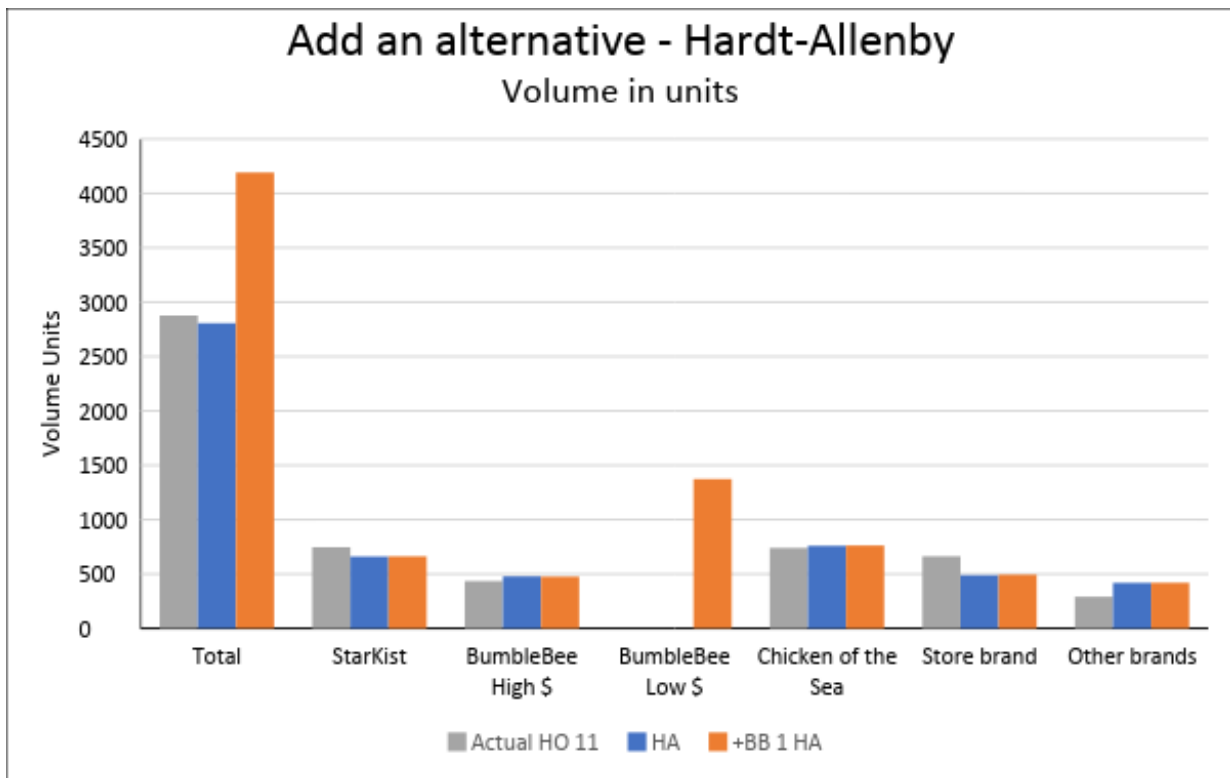
The results look like what we would expect when adding an alternative. The JDC model (Figure 12) shows an increase in total task volume. The original brands of 12 oz cans show a decrease in the number of units being sold, while the new SKU captures some of their volume and some latent demand (for 6 oz cans).

Figure 12. Volume in Units for Adding a Single Alternative—JDC Model



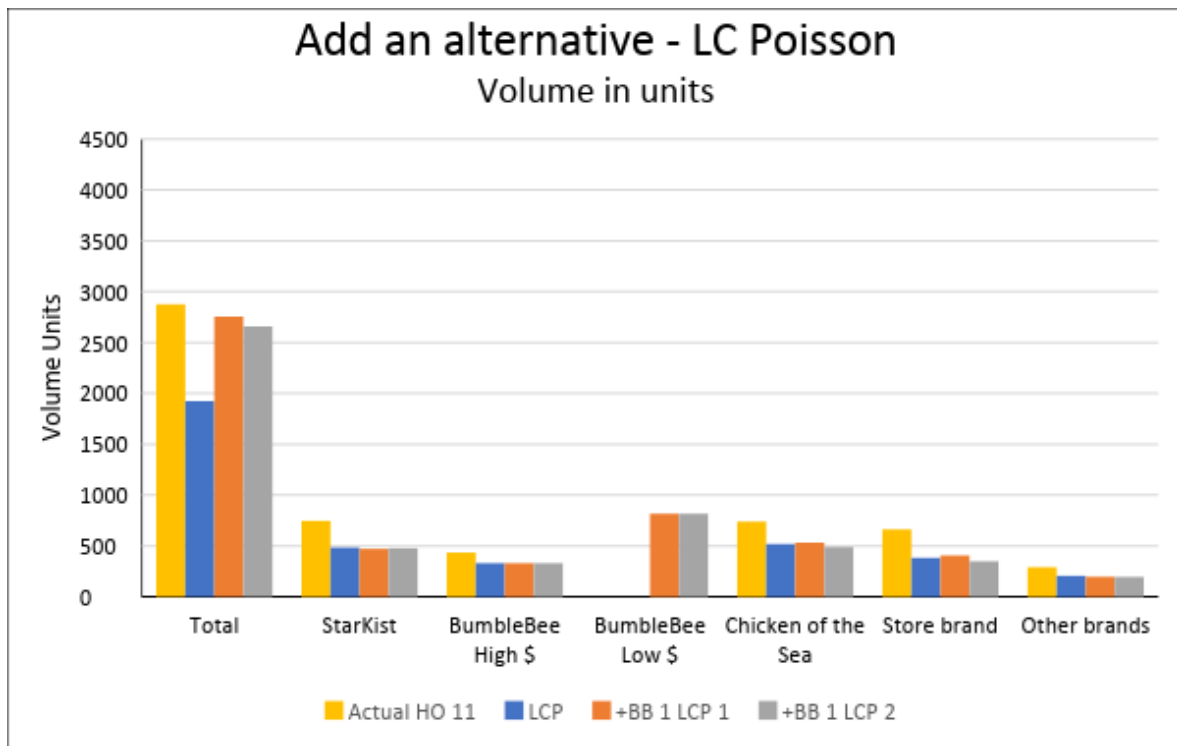
The Hardt-Allenby model (Figure 13) suggests the volume in units for the new Bumblebee SKU is from pure latent demand. Like the previous comparison, there is no meaningful change in the volume for the original alternatives. The change in unit volume for the 5 original brands ranged from -3 to +6 units on an average of 563 aggregate units. This suggests all new volume came from outside the original 5 brand market. As a result, the total task volume is all from latent demand. The increase in total task volume is much larger than the other models.

Figure 13. Volume in Units for Adding a Single Alternative—Hardt-Allenby Model



The chart from the LC Poisson model (Figure 14) is different from the other models. Note there are two predictions for the LC Poisson model: “+BB 1 LCP 1” and “+BB 1 LCP2”. These are two different ways to attempt to predict using a cross-effects model with a fixed number of brands in a task. The results associated with “+BB 1 LCP 1” were derived by duplicating the BumbleBee’s cross effects (and parameters) in the other alternatives’ model predictions. That is, a single model for each of the other brands attempt to capture both BumbleBee SKUs within their prediction. The original BumbleBee equations were used to predict the volume going to each SKU using the appropriate attribute levels for each BumbleBee SKU. The second set of predictions took the approach of predicting two sets of results for every brand and SKU using the appropriate levels of the existing and new BumbleBee SKUs for the cross effects. The predictions for each pair of equations were averaged to produce the results on the chart. While there are minor differences in the predictions for the Chicken of the Sea brand than the Store brand, the other alternatives share similar results across the two methods.

Figure 14. Volume in Units for Adding a Single Alternative—LC Poisson Model

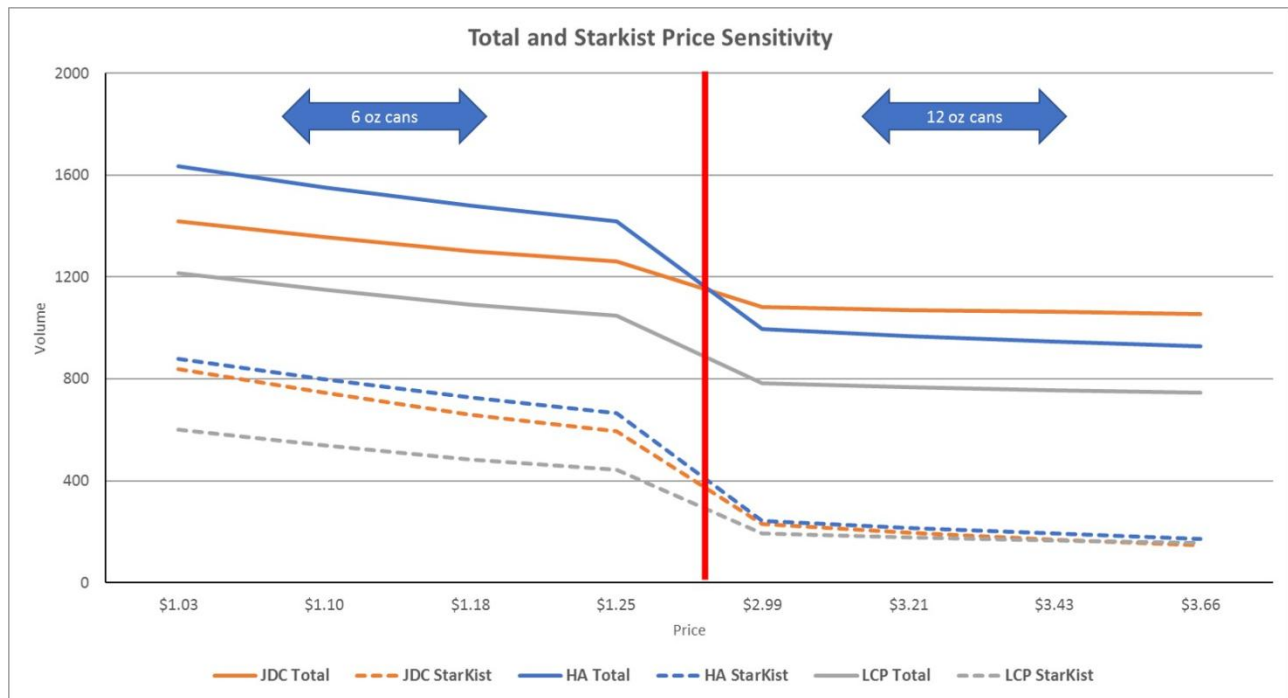


The results of the LC poisson model resemble the Hardt-Allenby model. There is very little change in the unit volume for the original brands. Almost all the volume associated with the new BumbleBee SKU is generated from outside the original market—that is, latent demand. The predicted volume for the Store Brand and the Chicken of the Sea using the “+BB 1 LCP 1” method are an anomaly in that they show an increase in volume when the new BumbleBee SKU is added. This is likely the result of overfitting with the cross effects and the duplication of the cross-effect parameters in this approach. The issue of handling the addition of new alternatives to cross-effect models is an area where future research needs to be done.

Price Sensitivity of a Brand in a Market

In this comparison we go back to using holdout task 9. We examine the price sensitivity of a single alternative across its entire breadth of prices (i.e., across both 6 oz and 12 oz cans). We arbitrarily chose to change the price of the StarKist brand. The attributes and prices of the other brands are held constant. Model predictions are used to build a set of price sensitivity curves. We show what happens to the total predicted task volumes, the volume of StarKist, and that of the other brands as we move from the lowest 6 oz can price to the highest 12 oz can price for StarKist. Figure 15 shows the prediction for the total volume and StarKist-own volumes as we change StarKist’s prices. Figure 16 shows what happens to the volumes of all the brands when StarKist changes its prices.

Figure 15. Volume in Units for Total Task Volume and StarKist's Own Volume When StarKist Prices Change



The three upper lines in Figure 15 show the change in total task volume as StarKist's prices increase from \$1.03 for a 6 oz can to \$3.66 for a 12 oz can. The vertical line and clear drop in volume in the price curves indicate where the StarKist SKU changed from the 6 oz to 12 oz can. There is a clear preference for the 6 oz can, especially when one can buy two 6 oz cans for less than a single 12 oz can. This pricing is clearly a design flaw. Except for a promotion, it is unlikely one would ever see 6 oz cans priced at less than 50% of the 12 oz can. However, given that a StarKist 6 oz can was never shown side-by-side with a 12 oz can, this pricing dominance was likely not apparent to respondents.

The total volume curves behave as one might expect. The three solid, upper curves depict each model's total task volume predictions. They reveal a similar pattern: dropping volume as price increases. The JDC model shows the least amount of absolute unit change from the lowest to highest price. The Hardt-Allenby shows the greatest absolute change in volume. The Hardt-Allenby model predicts the greatest volumes among the 6 oz cans, whereas the JDC model predicts the highest volume among the 12 oz cans. This is likely due to the strong reliance of the Hardt-Allenby model on price and the budget constraint. Recall, in the Hardt-Allenby model price is not a traditional attribute. All the non-price attributes utilities are directly scaled in price units. The other two models impose no budget constraint and price is treated a typical attribute.

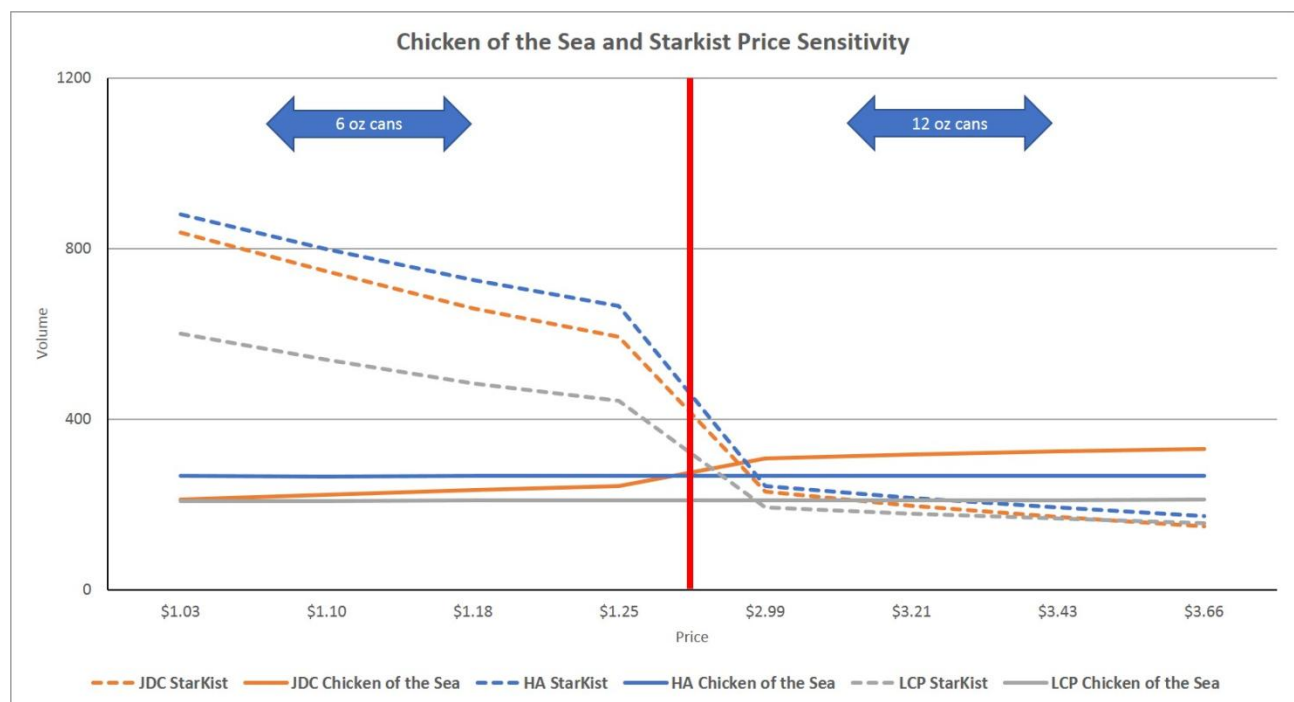
The StarKist-own (dashed) price curves reveal a similar pattern: as we increase price, the StarKist volume goes down—with the discontinuity between the 6 and 12 oz cans also present. The largest difference between the three models is in the 6 oz can price range. For the 12 oz cans they are similar.

When we begin to examine the impact of StarKist price changes on the other alternatives (Figure 16) we really see a difference among the models. Figure 16 shows the impact of changing

StarKist prices on Chicken of the Sea. The StarKist-own price sensitivity is shown, again, as dashed curves. They are the same as Figure 15. The solid curves near the bottom of the chart reveal the changes in the volume of Chicken of the Sea that occur when StarKist changes its prices and shifts from a 6 oz to a 12 oz can. The curves of the other brands (not shown) behave the same as that of Chicken of the Sea.

In Figure 16 we see as StarKist prices rise, the JDC model predicted volume of StarKist drops and predicted volume of Chicken of the Sea rises, albeit slowly. This aligns with what we might expect given the nature of the JDC model. The Hardt-Allenby model and the LC Poisson model depict curves that are virtually flat. The predictions show a very slight increase in Chicken of the Sea volume for the LC Poisson model, but the Hardt-Allenby model predictions are truly flat. As StarKist raises its prices, the volume held by StarKist in the Hardt-Allenby model is lost completely. The volume leaves the market. This is contrary to basic market expectations in a market of substitutable goods. Aggregated actual sales purchase data suggest some substitution in the real world, when prices change and/or promotions occur. The Hardt-Allenby model results are unreasonable if clients were expecting to see substitution among the products and they see all volume change NOT coming from other alternatives.

Figure 16. Volume in Units All Brands When StarKist's Prices Change



Add an Additional SKU to Each Brand in a Market

In this comparison we use holdout task 11 again and we add an additional SKU to every brand. In this case we added a 6 oz can for every brand. The other non-price attributes are held the same across all alternatives. We set price for the 12 oz cans and the 6 oz cans to each brand's respective price mid-point (i.e., half way between the highest and lowest prices for the respective can sizes). Rather than showing all the charts again, we summarize the results.

Basically, the JDC model behaves as before. There is an increase in total task volume and there is substitution among the brands. The total task volume increases by ~41%. The original alternatives (12 oz cans) all lost some volume roughly proportional to the IIA assumption of the MNL component of the JDC model (the drop ranged from relative amounts between -18% to -25.1%). For the total task volume, the increase was 702 units. The original brands lost 342 units (across all 5 brands) and the new brands gained 1,044 units ($702 = 1,044 - 342$). This appears consistent with our original substitution expectation.

The Hardt-Allenby model behaved as the previous comparisons. Total volume rose by 92.1% (2,158 units!). The original 5 brands lost a total of 2 units across all 5 brands (ranging from -4 to + 3 units). All the increase in total task volume originates from latent demand uncovered by adding the 5 new 6 oz cans to the task. The model predictions suggest, again, there is no substitution at all among the SKUs.

The LC Poisson model fails badly in this comparison. We only tested the adding of alternatives to the scenario and averaging across them to test the LC Poisson model in this case (i.e., we used the “+BB1 LCP 2” approach described earlier). The increase in total task volume is 116.2% (1,976 units!). Some of the existing alternatives gained volume (i.e., BumbleBee +13.5%; Chicken of the Sea +37.8% and other brands +34.9%). The other brands lost volume as we might expect (StarKist -11.8%; Store brand -8.9%). Of course, all the new 6 oz cans gained volume. These results would suggest a strong market complementarity which is unrealistic in this market. The mix of existing alternatives gaining and losing volume is really an example of cross effects having overfit the model. This comparison really stretches the capabilities of any cross-effects model and demonstrates the weakness of such models in predicting beyond the domain in which they were estimated.

CONCLUSION AND DISCUSSION

Volumetric modeling is hard. It is in its infancy in the field of marketing. There is disagreement on the best way to model volume. The disagreements focus on the differences between models of prediction and explanation, they concern whether a respondent has the ability answer volumetric tasks, and the different issues faced by practitioners vs. academic modelers. Regardless of what side you are on in any of these issues, volumetric modeling should no longer be avoided.

There are several points we wish to raise with these data and model predictions. They include recommendations on the design of volumetric tasks, and a discussion of the model results and some empirical evidence of support to the Hardt-Allenby predictions.

Survey Design Recommendations

The design used in this study has some weaknesses. Namely, the design used did not provide any data that would allow us to support or refute any of the predictions of the comparisons where alternatives were added or removed. The tasks always consisted of a single SKU from each brand. The SKUs varied, but there were never two SKUs from a single brand presented in a task together. Additionally, no brands were dropped from a task. Because of these design characteristics, we have no data to refute the results of the Hardt-Allenby model, or support that of the joint discrete/continuous model when adding or dropping alternatives. The patterns of substitution we see from the model predictions are based solely upon the assumptions inherent in the models. In the case of the joint discrete/continuous model, substitution is “forced” by the MNL component of the

model. The lack of substitution in the Hardt-Allenby model may be a function of the data or the model's method of capturing substitution through the outside good or a lack of identification in some of the key parameters.

Our recommendations for building volumetric designs are the following:

1. Build designs that enable the measurement of within-brand substitution—have more than 1 SKU per brand where possible.
2. Use a presence/absence design to control the presentation of the SKUs of interest—especially if absence on the shelf is of interest.
3. Build designs using alternative-specific attributes—it is best to estimate alternative-specific parameters where appropriate. It is always easy to make alternative-specific attributes generic rather than the other way around.
4. Ensure there are many exposures to each level of each attribute.
5. Carefully build an appropriate context for the task. In these data the context for the task was the next shopping trip. In this context it is easy for the respondent to give a zero volume because that may have suggested they would wait until their next shopping trip to see a better set of alternatives and prices. A longer context, such as having the respondent consider the shopping behaviors over the course of a month could lead to quite different volumes and alternative choosing behaviors.
6. Examine your data carefully and look for outliers—both within each respondent and across respondents. We were less concerned with such outliers here because we limited the volumes to be within a range. One should cap or remove extreme outliers when open-ended volume responses are elicited.
7. Lastly, clean, and reclean, the data. Look not only for outliers, but also for inconsistent total task volumes. Someone who has zero total task volume in their last N tasks may be a quitter. Someone who sped through the task and have zero variance in total task volume may not be a good respondent.

Model Results

Looking at the holdout data, the Hardt-Allenby model is the winner, with the joint discrete/continuous model coming in second. The LC Poisson using cross effects is the third runner-up (though an HB Poisson would have done better . . .).

The Hardt-Allenby model is the clear exception with respect to its predictions and substitution patterns. It showed no substitution whatsoever. Is the lack of substitution a model flaw? Is it an estimation flaw (e.g., an improper specification of the Priors as Jake Lee's paper suggests, 2018)? Is it a design flaw, or is it a true representation of the data? That is hard to say with certainty.

The joint discrete/continuous model is a descriptive, predictive model of volume. It works well in markets where substitution is expected. It is flexible because any form of a choice model may be used to capture substitution. It is also flexible in that the volumetric component can be any form of count model: linear, Poisson, negative binomial, zero-inflated. It can contain bias adjustments or other terms that may directly affect volume. And, using a system of models that nest alternatives, one can even capture complementarity across the nests of alternatives using cross effects in the volume component. It may not fit our data as well as the Hardt-Allenby model, but it will always work when substitution among products is expected.

The LC Poisson model, as we specified it, should never be used. Separate count models using cross effects with no constraints on the nature of those cross effects or across the system of equations is not an appropriate approach to modeling demand. The approach taken in the paper was naïve and the model clearly overfitted. Would a hierarchical Bayesian model have done better? In terms of overall predictability, yes. But, without constraints on the magnitude and sign of the cross effects, severe overfitting will occur. The impact of the cross effects would not make sense for many of the respondents and likely the sample as whole.

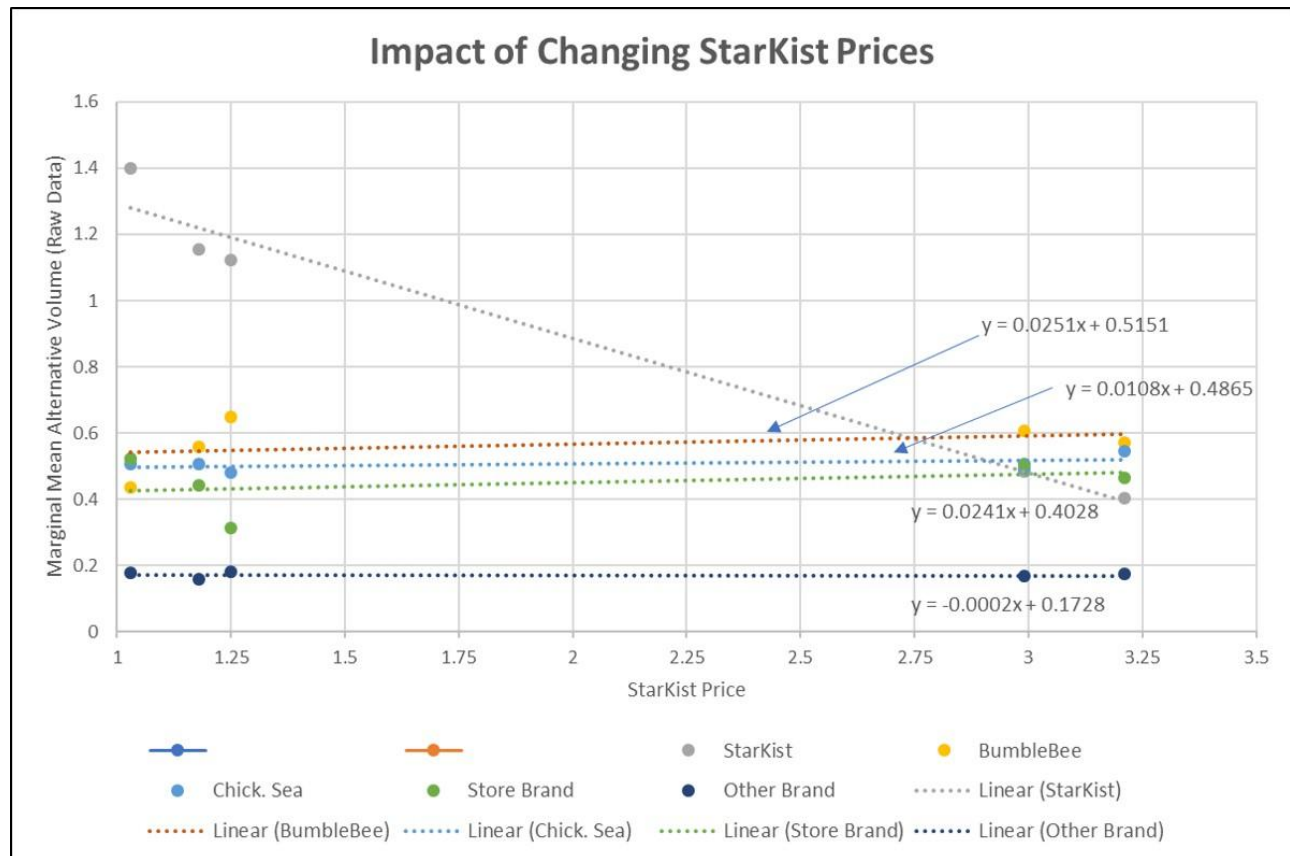
Can the Hardt-Allenby model capture substitution? Professor Nino Hardt ran some simulations using their model when we shared our results with him prior to presenting the paper. In those simulations he found that the model's satiation and budget parameters affect predictions of substitution in their model. The simulations showed that reducing the magnitude of the satiation and budget parameters (i.e., bringing them closer to zero), increases the amount of substitution among alternatives in a task *ceteris paribus*. The Hardt-Allenby model in this paper has a budget parameter that is higher and more highly variable than one might typically expect².

Our discussant, Greg Allenby (2018), in reviewing the presentation, suggested there might also be a “siloeing” effect in our data. That is, respondents' stated volumes stayed within a specific attribute level, or levels, and the respondents never varied this behavior across tasks. We call this a repertoire effect. For example, one might have always assigned a nonzero volume to 6 oz cans from BumbleBee only, and, when that alternative did not appear, the total task volume went to zero. This is hard to examine within the context of a fractional factorial design. It might be easier to examine in a design where there are duplicated alternatives on all but one attribute (but such may be a bad design).

To test Greg's hypothesis, we produced a chart showing the mean alternative-specific volumes across tasks for all brands as StarKist's prices changed. These are not predictions but means across tasks of the raw data. There are only about 4-5 points per brand. The actual means are shown as different points in the chart. We allowed Excel to draw the regression line among each set of points to see what the slope of each line might be. If substitution occurs, we would expect the slope of each non-StarKist line to be positive and greater than zero. Figure 17 shows these results.

² The mean, median, and 95% confidence intervals for the Hardt-Allenby satiation and budget parameters are:
Satiation upper-level model parameter: mean = -0.900; median = -0.901; 5% = -0.843; and 95% = -0.957
Budget upper-level model parameter: mean = 3.270 (!); median = 3.269; 5% = 1.397; and 95% = 5.391
The budget parameter is highly uncertain!

Figure 17. Plot of the Marginal Means of the Raw Data Averaged Across Attributes for Each Brand While Changing the Available Prices of StarKist



This chart (Figure 17) shows clearly that the impact of raising StarKist’s prices on the average volume of the other brands is quite small. Except for Other brands, the slopes are all positive as we might expect if substitution were occurring, but they are all near zero. They are almost flat. This lends support to the results of Hardt-Allenby model. All price curves for the non-StarKist brands show the same pattern.

Nevertheless, one could argue that these results might be a difficult sell to managers who expect substitution in the market. We still are skeptical of the ability of Hardt-Allenby model to predict substitution. While this data seems to support the total lack of substitution among these alternatives, one still questions how well the Hardt-Allenby model would capture substitution in a data set where it is more pronounced. More importantly, as a practitioner, we are aware of our clients’ expectations. Many of us have been put into situations where we sometimes must sacrifice accuracy to present results that meet our client’s “gut feel”. In situations where substitution is expected, there is still some uncertainty of how well the Hardt-Allenby model would be accepted vis-à-vis a descriptive model such as the joint discrete/continuous model.

ACKNOWLEDGMENT

This work was partially supported by Social Sciences and Humanities Research Council (SSHRC) of Canada Grant (No. 430199).



Thomas C. Eagle



Jordan Louviere



Towhidul Islam

REFERENCES

- Allenby, Greg (2018), discussion of the presentation “A Comparison of Volumetric Models,” presented at the Sawtooth Software Conference
- Eagle, Thomas C. (2010), Modeling Demand Using Simple Approaches, Sawtooth Software Conference
- Garratt, Mark and Thomas C. Eagle (2010), Practical Approaches to Modeling Demand, ART Forum
- Hanemann, W.M. (1984), Discrete/Continuous Models of Consumer Demand, *Econometrica*, 52, 541–61
- Hardt, Nino, Youngju Kim, Mingyo Joo, Jaehwan Kim, and Greg Allenby (2017), Reconciling Stated and Revealed Preference, ART Forum.
- Hausman, J.A., G.K. Leonard, and D. McFadden (1995), A Utility-Consistent, Combined Discrete Choice and Count Model Assessing Recreational Use Losses Due to Natural Resource Damage, *Journal of Public Economics*, 56, 1–30
- Howell, John and Greg Allenby (2012), Choice Models with Fixed Costs, Sawtooth Software Conference
- Huber, Joel (2018), discussion of the presentation “Properties of Direct Utility Models for Volumetric Conjoint,” presented at the Sawtooth Software Conference
- Kim, Jaehwan, Greg Allenby and Peter Rossi (2007), “Product Attributes and Models of Multiple Discreteness,” *Journal of Econometrics*, 138, 208–230.
- Lee, Jake (2018), Properties of Direct Utility Models for Volumetric Conjoint, paper presented at the Sawtooth Software Conference
- Pachali, Max, Peter Kurz, and Thomas Otter (2017), The perils of Ignoring the Budget Constraint in Single-Unit Demand, working paper.
- Train, Ken (1986), *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*, Cambridge, MA: The MIT Press.

DIRECT ESTIMATION OF KEY DRIVERS FROM A FITTED BAYESIAN NETWORK

BENJAMIN CORTESE

KS&R

ABSTRACT

Key driver analysis (KDA) is a technique to identify a subset of attributes, known as key drivers, which have a strong impact on a target attribute, such as satisfaction, likelihood to purchase, or likelihood to recommend. There is a wide array of techniques to estimate attribute level driver scores, but those most commonly used are unable to provide information about the interactions between drivers. The introduction of Bayesian networks (BNs)—graphical representations of attribute relationships—help make sense of these complex interactions. Attempts to combine KDA and BNs through separate analysis often lead to conflicting results from the estimated top drivers and the attribute relationships depicted by the network.

We propose a new algorithm, BNKDA, to calculate driver scores directly from a fitted Bayesian network. This method relies on the Max-Min Hill-Climbing (MMHC) network fitting algorithm, Bayesian Information Criterion (BIC) and arc strengths calculated from the network. A weight factor is suggested for use to reduce the impact of longer paths to the target attribute. This technique provides both the directed acyclic graph (DAG) visualizing attribute relationships and corresponding driver scores to tell a cohesive story.

The algorithm is compared to two widely adopted driver analysis methods—Kruskal’s relative importance (a variant of a Shapley value) and partial least squares path modeling (PLSPM)—through simulation studies. It is found that all three techniques identify similar top drivers in terms of ordering, but the magnitude of scores differs. The regression based methods (Kruskal and PLSPM) favor directly impacting attributes in the hierarchy, while BNKDA provides more balanced estimates. Consistency of driver estimates obtained from BNKDA imply that this is a viable option to calculate driver scores directly from a BN.

INTRODUCTION

Key driver analysis (KDA), also known as relative importance, is a longstanding market research method to identify the strength of relationships of a set of attributes and their influence on performance metrics such as satisfaction or customer loyalty. This is an exploratory technique, providing hypotheses about possible relationships that are used to guide recommendations that will most likely lead to improvement of the target metric.

As an example, consider a business that has been tasked with improving satisfaction with a specific product. In order to pinpoint a few key areas to focus resources, a survey was conducted to measure consumer ratings in areas such as product warranty, ease of use, product value, customer service, etc. The raw ratings data provides little differentiation between attributes, making it difficult to suggest strategic recommendations. The typical solution to this problem is KDA, run with product satisfaction as the target attribute and the others as predictors. A driver score (also referred to as an index) is assigned to each predictor to indicate the level of

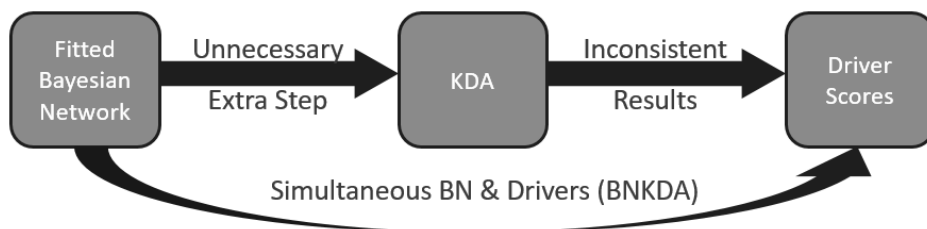
importance or impact that attribute has on satisfaction. The attributes with larger scores are identified as possible drivers, and can then be targeted with additional resources to improve overall satisfaction.

There are many approaches to estimate driver scores and those most commonly applied are regression based methods. These techniques include correlation analysis, multiple regression and Kruskal's relative importance, and are used to provide a single importance score associated with each attribute. Correlation analysis is a very simple technique with fast turnaround time and does not require a large sample. Multiple regression is similar to correlation but requires model validation and can be misleading in the presence of multicollinearity. Kruskal's relative importance, a variant on the Shapley value for R-squared, accounts for multicollinearity by averaging over each squared partial correlation for all predictor attributes, but does not support predictive follow-up analysis (Kruskal, 1987). These methods are designed to pinpoint the most relevant drivers but are not capable of modeling the interrelationships between attributes.

The growing popularity of Bayesian networks (BNs) has opened the door to innovative techniques that provide information on the entire ecosystem of attributes, as well as the traditional attribute importance score. The BN methodology has been adopted over the past several years, but the direct calculation of the importance score from a BN remains to be explored.

At first, a two-step process was developed to generate importance scores using the BN. The network structure, i.e., all information regarding attribute relationships, was obtained by fitting a BN to the data. The importance scores were then estimated using a subsequent analysis of the network structure, such as partial least squares path modeling (PLSPM), structural equation modeling (SEM), or probabilistic structural equation modeling (PSEM) (Vinzi, Trinchera, & Amato, 2010). Specifically, the scores were calculated by taking the product of the outer weights and total effects to the target attribute. This method is referred to as BNPLSPM. While it has proved sufficient, BNPLSPM can sometimes lead to inconsistencies between the network model and the importance scores, such as overstatement of direct drivers or understatement of indirect drivers.

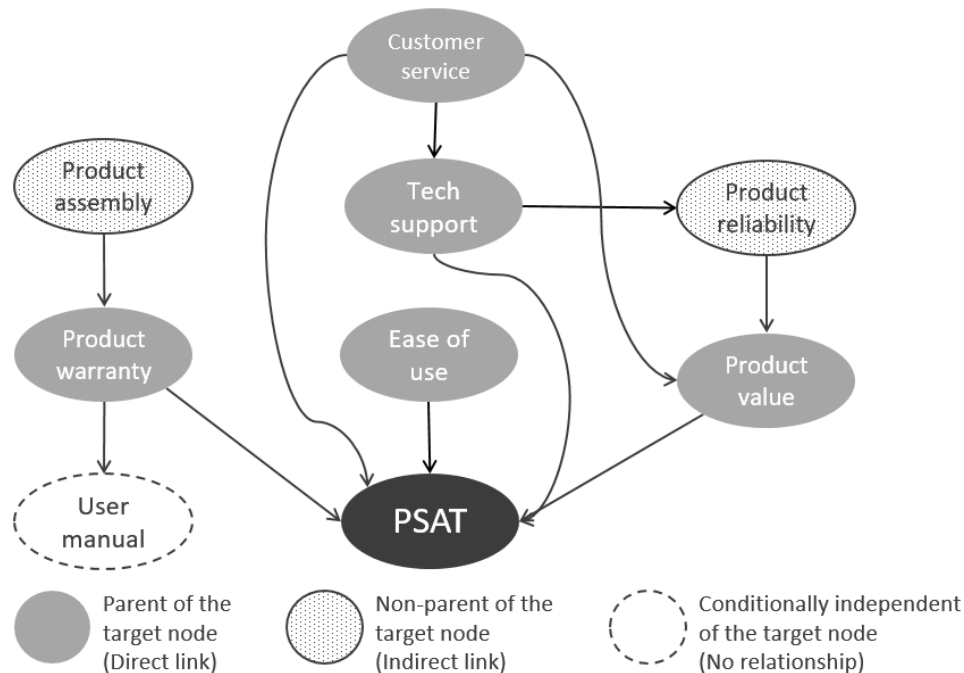
These inconsistencies identified the need for a different approach. This paper introduces an algorithm, Bayesian Network Key Driver Analysis (BNKDA), to calculate importance scores directly from a BN, eliminating the need for subsequent modeling. No benchmark comparison method exists for KDA so instead, simulation results are used to compare the algorithm to the widely accepted Kruskal's relative importance index and two-step procedure BNPLSPM via consistency and estimates of driver scores.



BACKGROUND

A BN is a directed acyclic graph, or DAG, that models probabilistic dependencies and independencies among attributes, also known as nodes. These specific DAGs are graphical representations of conditional probabilities represented by arcs with arrowheads from the one node to another. Each arc must have a unique direction that implies a causal relationship with no loops in the graph. An example of a BN is given in Figure 1.

Figure 1. Sample BN



Each network is comprised of parent and child nodes, and occasionally independent or standalone nodes. For the following definitions, see Figure 1 for reference.

- Parent node: A node with at least one outgoing arc to another node in the BN, such as *Customer service*.
- Child node: A node with at least one incoming arc from another node in the BN, such as *User manual*.
- Independent node: A node with no path to the target attribute, such as *User manual*.
- Terminal node: A node with no children, such as *User manual*.
- Initial node: A node with no parents, such as *Customer service*.

A target attribute is required when fitting any driver analysis model, but is not necessary when fitting a BN. Because of this, the target attribute (or attributes) must be specified prior to modeling, forcing this attribute to be a terminal node. Other terminal nodes may be present in the final fitted network, for example, *User manual* in Figure 1. This indicates that *User manual* is conditionally independent of PSAT.

Fitting the BN

The proposed algorithm requires a fitted BN. While there are a wide variety of fitting algorithms, the algorithm of choice for this paper is Max-Min Hill-Climbing (MMHC) (Tsamardinos, Brown, & Aliferis, 2006). For information on other fitting algorithms, see (Daly & Shen, 2007; Margaritis, 2003; Russell, 2009; Tsamardinos, Aliferis, & Statnikov, 2003).

The MMHC algorithm fits the network in a two stage process.

First, in the Max-Min portion of the algorithm, candidate sets of all possible parent and child nodes are identified through tests of conditional independence individually, for each attribute in the data. At this stage, related attributes are grouped into sets, but no orientation is assigned.

Then, starting from an empty graph, a Greedy Hill-Climbing search is performed by iterating through edge operations, such as edge addition, deletion, and orientation reversal, to incrementally improve the BIC. In this particular case, “Greedy” indicates that the only edges included in the Hill-Climbing stage are those identified as possible parents or children in the Max-Min stage. The search returns the BN (with orientation) that scored the optimal BIC.

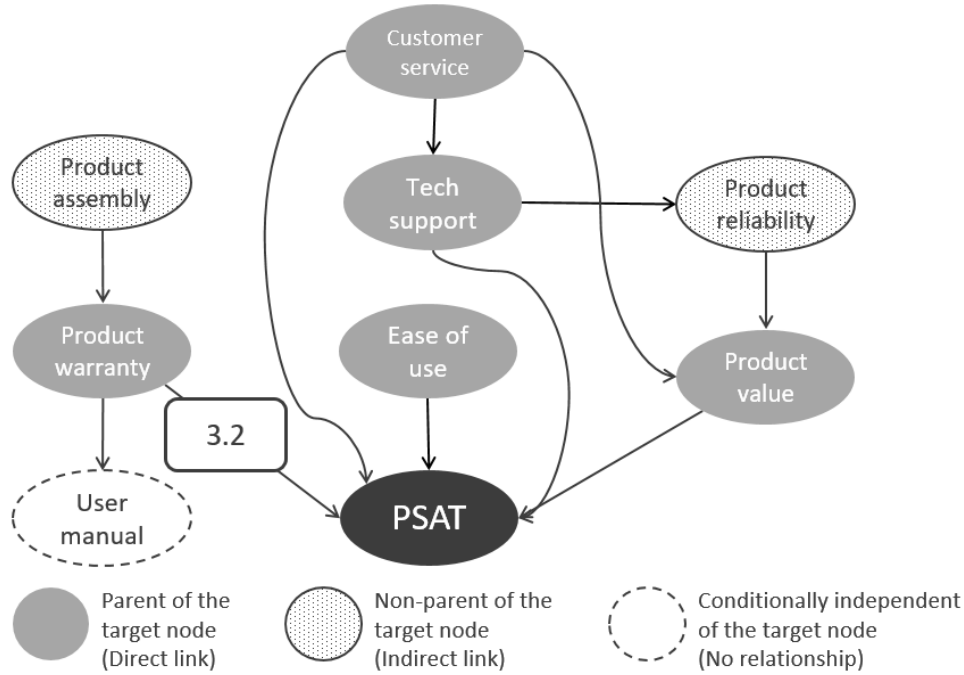
This MMHC fitting algorithm is currently one of the most popular used in practice due both to its efficiency and accuracy.

Arc Strength

Once the structure of the BN is fit using MMHC, the strengths of the relationships between attributes, known as arc strengths, are estimated. There are various methods to estimate arc strength, such as linear correlation, mutual information, contribution to AIC or BIC, and others specific to networks estimated from discrete data (Ebert-Uphoff, 2007; Nicholson, 1998). For the purposes of this paper, the arc strength is calculated via the contribution to the BIC.

This method quantifies every arc in the network, assigning a value denoted by $STR(Node_1 \rightarrow Node_2)$, and is based on that arcs contribution to the overall BIC of the network. This contribution is calculated by removing each arc from the network, one at a time, and recalculating the BIC of the resulting simpler network.

Figure 2. Sample BN with Arc Strength



As an example, consider the BN in Figure 2, with arc strength for *Product warranty* → *PSAT* listed as 3.2. Suppose that the BIC of the full network is 100. When that arc is removed, the BIC of the simplified network without that arc is 96.8, a decrease of 3.2, defining:

$$STR(Product\ warranty \rightarrow PSAT) = 3.2.$$

Note that the definition of the BIC is rescaled by a factor of -2, so in this case, the higher the BIC, the better the network. All arc strengths of this network are shown in Table 1 for reference.

Weight Factor

The purpose of any KDA exercise is to provide strategic recommendations on where to focus resources from a typically large set of attributes. This decision relies on the methodology to pinpoint a smaller subset of highly impactful attributes on the target metric.

Typical BNs are complex, with some attributes having dozens of paths to a target attribute. These highly connected attributes are usually non-parents of the target, with longer paths to reach the target. This raised the question of how to handle the proximity, or path length, from an attribute to the target.

Any parent of the target attribute should have the opportunity to have a larger driver score than non-parents. Because of this, a weight factor is introduced (denoted $w(l)$), prioritizing proximity to the target by penalizing attributes with longer paths to reach the target. The purpose of the weight factor is to balance connectivity (many paths) and proximity (shorter path length) to the target attribute.

Four path weights are introduced— $w(l) = 1, \sqrt{l}, l$ and $l!$ —and their impacts explored via simulation.

BAYESIAN NETWORK KEY DRIVER ANALYSIS

The BNKDA algorithm described requires a fitted Bayesian network with estimated arc strengths, and produces an importance score for each attribute in the model. For the purposes of this paper, the network is fit using the MMHC algorithm, arc strengths are produced via the contribution to the BIC method and the scores are calculated based on a summary function, typically the sum, which summarizes the strengths of all complete paths from each driver to the target.

The Algorithm

The first step of BNKDA is to clean the network by removing any nodes that are conditionally independent of the target. This includes all (non-target) terminal nodes and nodes whose children are only (non-target) terminal nodes, among others. For example, in Figure 2, *User manual* will be removed.

After the network is cleaned, the arc strengths are standardized. The standardization procedure is intended to prevent a handful of large arc strengths from dominating the driver score estimates. For example, the arc from *Customer service* to *Tech support* in the sample BN has a strength of 191.8, while the largest arc strength of a parent of *PSAT* is 59.3. The arc strength of 191.8 is obviously an important connection, but its relevance to the driver score estimates of *PSAT* is limited. All standardized arc strengths for the sample network are shown in Table 1 below, for reference.

The standardization procedure works in the following way. Starting with the target node, the arc strengths of all arcs originating from the parent nodes are standardized relative to one another. That is, each arc strength from the parent node to the target node is divided by the sum of the arc strengths of all parents of the target. The result is a set of arc strengths ranging from 0 to 1 where the stronger relationships are rewarded with larger values.

The standardization procedure then shifts to the next level of the network and repeats, standardizing all arcs of corresponding parent nodes for each child node in the network. If a child node has a single parent, that arc strength is standardized to 1 regardless of the other arc strengths in the network.

Standardized arc strengths are formally defined as follows. Let N_{c_i} denote the i^{th} child node in the network with n_{c_i} total parents and $P_{c_i,1}, P_{c_i,2}, \dots, P_{c_i,n_{c_i}}$ denote the parents of N_{c_i} . For fixed i and for all $j = 1, 2, \dots, n_{c_i}$, the standardized arc strength for the arc $P_{c_i,j} \rightarrow N_{c_i}$ is:

$$STR_{STD}(P_{c_i,j} \rightarrow N_{c_i}) = \frac{STR(P_{c_i,j} \rightarrow N_{c_i})}{\sum_k STR(P_{c_i,k} \rightarrow N_{c_i})}.$$

Table 1. Arc Strengths of the Sample BN

Parent	Child	Raw strength	Standardized strength
Ease of use	PSAT	59.3	0.41
Product value	PSAT	56.2	0.39
Tech support	PSAT	14.4	0.10
Customer service	PSAT	12.3	0.08
Product warranty	PSAT	3.2	0.02
Product reliability	Product value	1.7	0.17
Customer service	Product value	8.3	0.83
Customer service	Tech support	191.8	1.00
Product assembly	Product warranty	0.7	1.00
Tech support	Product reliability	59.5	1.00

Determining the strength of the relationship of a predictor node to the target requires estimation of the strength of every possible arc to that node. In Figure 1, *Ease of use* has a single path to the target node:

$$Ease\ of\ use \rightarrow PSAT,$$

whereas *Customer service* has four paths:

$$Customer\ service \rightarrow PSAT,$$

$$Customer\ service \rightarrow Product\ value \rightarrow PSAT,$$

$$Customer\ service \rightarrow Tech\ support \rightarrow PSAT,$$

$$Customer\ service \rightarrow Tech\ support \rightarrow Product\ reliability \rightarrow Product\ value \rightarrow PSAT.$$

The strength of each of these paths will be used to calculate the final node strength.

The raw path strength (RPS) of each path from one node to the target node is calculated by multiplying the standardized path strengths together. Define a generic path of length l as:

$$Path := Node_0 \rightarrow Node_1 \rightarrow \dots \rightarrow Node_l,$$

then the raw path strength is calculated as:

$$RPS(Path) = \prod_{i=0}^{l-1} STR_{STD}(Node_i \rightarrow Node_{i+1}).$$

Next, the weight factor $w(l)$ is applied to each RPS, resulting in a weighted path strength (WPS).

$$WPS(Path) = \frac{RPS(Path)}{w(l)}.$$

The idea of a path weight for modeling node importance was first defined for mutual information in section 5 of (Nicholson, 1998). The impact of the choice of weight factor is explored in the simulation study using four candidate weights—1, \sqrt{l} , l , and $l!$. Both the RPS and WPS calculated from the sample BN (Figure 1) are available in Table 2 below, with $w(l) = l!$.

Table 2. RPS and WPS of the Sample BN

Path	RPS	$w(l)$	WPS
<i>Customer service</i> → <i>Tech support</i> → <i>Product reliability</i> → <i>Product value</i> → <i>PSAT</i>	0.064	4!	0.003
<i>Customer service</i> → <i>Tech support</i> → <i>PSAT</i>	0.099	2!	0.049
<i>Customer service</i> → <i>Product value</i> → <i>PSAT</i>	0.322	2!	0.161
<i>Customer service</i> → <i>PSAT</i>	0.084	1	0.084
<i>Tech support</i> → <i>Product reliability</i> → <i>Product value</i> → <i>PSAT</i>	0.064	3!	0.011
<i>Tech support</i> → <i>PSAT</i>	0.099	1	0.099
<i>Product reliability</i> → <i>Product value</i> → <i>PSAT</i>	0.064	2!	0.032
<i>Product value</i> → <i>PSAT</i>	0.386	1	0.386
<i>Product assembly</i> → <i>Product warranty</i> → <i>PSAT</i>	0.022	2!	0.011
<i>Product warranty</i> → <i>PSAT</i>	0.022	1	0.022
<i>Ease of use</i> → <i>PSAT</i>	0.408	1	0.408

The weighted path strengths are combined into a raw node score (RNS) by summing all path strengths with common initial node, that is:

$$RNS(Node_i) = \sum WPS(\text{All Paths with Initial Node} = Node_i)$$

The WPS defines the importance index for each node from the fitted network. If desired, the scores can be standardized to sum to 1 for ease of interpretation. The scores are listed in Table 3 below.

Table 3. Raw and Standardized Node Scores from the Sample BN

Node	Raw Score	Standardized Score
Ease of use	0.408	0.322
Product value	0.386	0.305
Customer service	0.298	0.235
Tech support	0.110	0.086
Product reliability	0.032	0.025
Product warranty	0.022	0.017
Product assembly	0.011	0.009
User manual	0.000	0.000

Summary of BNKDA

A brief step-by-step summary of the Bayesian network key driver analysis algorithm is listed below.

1. Fit the Bayesian network using MMHC.
2. Calculate the arc strength for each arc in the network using the contribution to the BIC.
3. Standardize the arc strengths according to the corresponding parent nodes.
4. Calculate raw path strengths by multiplying the standardized arc strengths in each path.
5. Weight the raw path strengths via a weighting factor $w(l)$.
6. Calculate node strength by summing all path strengths with common initial node.
7. Report the BN importance index as the standardized node score.

SIMULATION RESULTS

A simulation study was designed with two goals in mind. First, to explore the estimated driver scores using BNKDA with a variety of data structures; and second, to better understand the impact of the weight factor $w(l)$. Three models were developed using sample sizes varying from 75 to 1,000 and are outlined in detail in the Appendix. The choice of this range is most representative of sample sizes common to market research, as samples of size greater than 1,000 are often out of the scope of a study.

It is well known that the ability to successfully fit the correct structure of a BN relies on a relatively large sample size. Small samples were included in the simulation study to compare results with Kruskal's relative importance, which works quite well in the presence of small samples.

For each of the three models and each sample size, 1,000 data sets were simulated using the statistical language R. Driver scores were modeled using both Kruskal's relative importance (from the *relaimpo* package) and BNKDA (fit using the *bnlearn* package) for each data set and averaged together to analyze overall performance and differences in method behavior (Grömping, 2006; Scutari, 2010). Comparison between BNPLSPM (using SmartPLS 3) and BNKDA is also provided, but due to time and resource restrictions, models are limited to a single data set for each simulation (Ringle, Wende, & Becker, 2015).

BNKDA Performance and Driver Estimation

Each of the simulated data sets was designed to study specific aspects of the BNKDA algorithm and associated weight factor. In this section, all reported statistics are computed from 1,000 randomly generated data sets, each with a sample of size 500. The sample of size 500 was selected based on typical sample sizes present in market research studies.

The intended structure for each BN is provided along with the top four drivers (after rescaling and taking averages) for each weight factor for Models 1 and 2, and the entire table for Model 3. The columns of each driver table are listed in decreasing order of penalty on longer paths, starting with the largest weight, the factorial.

Model 1

The data representing Model 1 is intended to simulate a typical driver analysis model with eight nodes, one of which is conditionally independent of the target, and a two level hierarchy.

Figure 3. Model 1 Intended Network Structure

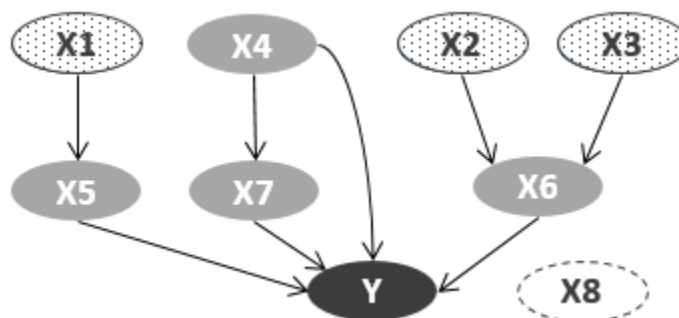


Table 4. Model 1 Rescaled Mean Driver Index for Top 4 Attributes (N = 500)

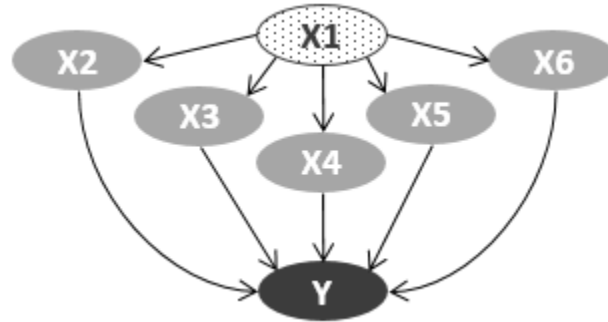
$w(l) = l!$	$w(l) = l$	$w(l) = \sqrt{l}$	$w(l) = 1$
X6 29.0%	X6 29.0%	X6 25.5%	X6 21.9%
X5 27.3%	X5 27.3%	X5 24.1%	X5 20.7%
X1 13.7%	X1 13.7%	X1 17.1%	X1 20.7%
X7 8.0%	X7 8.0%	X2 9.2%	X2 11.2%

The top three drivers are identical for each of the four weight factors, with the fourth changing from X2 to X7 for larger weights. This is likely because the smaller values of $w(l)$ enable longer paths to have larger strength (see Figure 3, nodes X1, X2 and X3) and, in turn, directly linking nodes will have lower relative driver scores in the presence of many indirect nodes.

Not shown in Table 4 are the driver scores for the conditionally independent node X8. All four weights resulted in an average standardized driver score of 0%, indicating that a sample size of 500 was adequate to detect the conditional independence in this particular model. (Samples as small as $N = 150$ produced an average driver score of 0% for X8.)

Model 2

The data for Model 2 is used to explore the impact of a highly connected node with no direct link to the target.

Figure 4. Model 2 Intended Network Structure**Table 5. Model 2 Rescaled Mean Driver Index for Top 4 Attributes (N = 500)**

$w(l) = l!$	$w(l) = l$	$w(l) = \sqrt{l}$	$w(l) = 1$
X4 39.8%	X4 39.8%	X1 40.4%	X1 48.7%
X1 32.5%	X1 32.6%	X4 35.2%	X4 30.2%
X6 10.3%	X6 10.3%	X6 9.2%	X6 7.9%
X2 7.7%	X2 7.7%	X2 6.8%	X2 5.9%

The results from Model 2 are precisely as expected. Node X1 in Figure 4 has no direct link to the target node, but has five paths through directly linking nodes. Because of this, X1 should be a top driver for any choice of weight and, as shown in Table 4, it is in the top two for all simulations. The larger choices of $w(l)$ allow a directly linking node, X4, to have the largest driver score, while the smaller choices of $w(l)$ lead to node X1 as the top driver. The above results indicate that at least some weight should be applied to each RPS to prevent a node such as X1 from dominating the output.

Model 3

Model 3 is intended to analyze the effect of differing path lengths on node importance. This model is made up of four single paths to the target of lengths one, two, three and four.

Figure 5. Model 3 Intended Network Structure

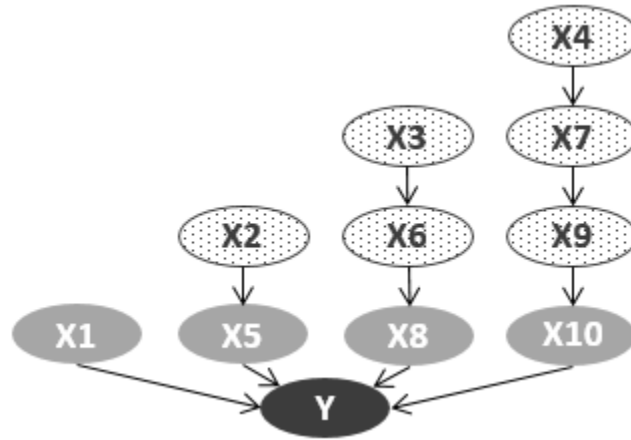


Table 6. Model 3 Rescaled Mean Driver Index for All Attributes, Top 4 Highlighted (N = 500)

Node	$w(l) = l!$	$w(l) = l$	$w(l) = \sqrt{l}$	$w(l) = 1$
X5	24.3%	22.5%	18.5%	14.5%
X8	22.4%	20.6%	16.9%	13.1%
X10	12.7%	11.5%	9.3%	7.1%
X2	11.8%	11.0%	12.8%	14.1%
X6	11.4%	10.5%	12.1%	13.2%
X9	6.9%	6.3%	7.1%	7.5%
X3	3.8%	6.9%	9.7%	13.0%
X1	3.6%	3.4%	2.9%	2.3%
X7	2.5%	4.2%	5.8%	7.6%
X4	0.7%	3.1%	4.9%	7.4%

When varying path lengths are present in the BN, as shown in Figure 5, the choice of weight factor has a strong influence on the ability of BNKDA to differentiate a few key drivers from the rest. In this case, each node has exactly one path to the target, forcing any indirect node to have a lower driver score than those it must pass through.

When $w(l) = 1$, there are exactly four unique driver scores—those corresponding to the directly impacting nodes. (In the results above, the scores of the node sets (X2, X5), (X3, X6, X8) and (X4, X7, X9, X10) will be equal when the network structure is specified correctly. Using 1,000 random draws occasionally will produce a misclassification of the network structure, causing the inconsistencies in Table 5.) These results provide evidence that some weight factor is necessary to avoid ambiguity in drivers and counterintuitive results.

Comparison to Other Methods

Driver analysis techniques have no true benchmark comparison method to identify accuracy of the estimated driver scores. Instead, the driver estimates of BNKDA are compared to those of Kruskal's relative importance and BNPLSPM to identify similarities and differences in outcomes for various data structures.

The simulation study to compare these methods relies on the same synthetic data outlined in the previous section, with sample sizes ranging from 75 to 1,000. Samples of size 1,000 were used when comparing BNKDA to BNPLSPM in order to ensure correct estimation of the Bayesian network structure.

Kruskal's Relative Importance versus BNKDA

Overall, the results of Kruskal's relative importance are similar to those of BNKDA. Comparing Table 7 to those in the previous section (Tables 4–6), it is clear that both methods identify similar attributes as top drivers in each of the models, although the associated scores differ substantially from model to model. In particular, the top two drivers in Models 1 and 3 and the top four drivers in Model 2 are the same, up to ordering.

Table 7. Kruskal's Rescaled Mean Driver Index for Top 4 Drivers and All Models (N = 500)

Model 1		Model 2		Model 3	
x6	27.4%	x4	37.3%	x5	30.1%
x5	23.5%	x6	15.4%	x8	23.0%
x7	18.6%	x1	13.5%	x10	15.1%
x4	15.9%	x2	13.5%	x1	8.7%

The choice of weight factor plays a role in BNKDA achieving similar results to Kruskal's relative importance. When $w(l) = ll$, the two methodologies show the most similar estimates. With this choice of weight factor, we have:

- Model 1: Identical top two drivers with similar score estimates
- Model 2: Identical top four drivers (up to ordering)
- Model 3: Identical top three drivers with similar score estimates

When smaller values of $w(l)$ are selected, there are substantial differences in attribute rank and score estimates.

The key difference between these two methods, regardless of choice of weight, is that Kruskal's relative importance tends to favor directly linking nodes, assigning them higher driver scores than BNKDA. As shown in Table 7, the top four drivers for both Model 1 and Model 3 using the estimates from Kruskal's relative importance are those nodes with a direct link to the target. When using the estimates from BNKDA, X4 in Model 1 has scores ranging from 7.5% to 8.7%, nearly half that of Kruskal's estimate and, as shown in Table 6, no score of X1 in Model 3 is larger than 3.6%.

Another difference between the two methods is the ability to differentiate attributes with no impact on the target via driver scores of 0. For example, X8 in Model 1 represents an attribute that is independent of the target, simulated with no statistical relationship to Y, and should have an estimated driver score of 0. While Kruskal's relative importance does assign a small score to

X8, results in Table 8 show that BNKDA provides a correct driver estimate for X8 for samples as small as $N = 150$.

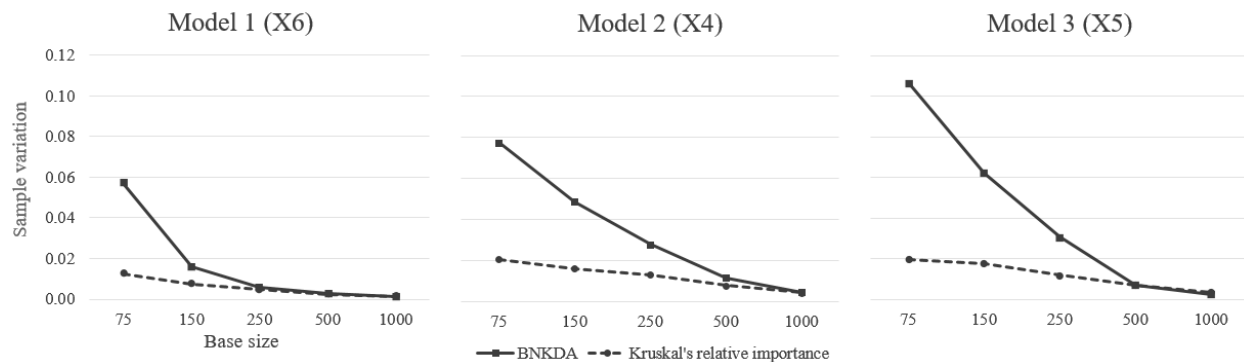
Table 8. Mean Standardized Driver Score for Independent Node X8 in Model 1

Sample size	Kruskal's relative importance	BNKDA ($w(l) = l!$)	BNKDA ($w(l) = l$)	BNKDA ($w(l) = \sqrt{l}$)	BNKDA ($w(l) = 1$)
75	2.5%	0.5%	0.5%	0.5%	0.5%
150	1.3%	0.0%	0.0%	0.0%	0.1%
250	0.8%	0.0%	0.0%	0.0%	0.0%
500	0.4%	0.0%	0.0%	0.0%	0.0%
1000	0.2%	0.0%	0.0%	0.0%	0.0%

Consistency of estimates is critical to any driver analysis technique. The variability of estimates for BNKDA is much larger than that of Kruskal's relative importance for small samples, but once modest sample sizes are achieved ($N \sim 500$), this disparity disappears. The small sample behavior is not a surprise, as the structure of a BN tends to be under-fit in the presence of small samples (Zuk, Margel, & Domany, 2012), while Kruskal's relative importance is not greatly impacted.

The sample variation of the top driver, as identified by BNKDA (with $w(l) = l!$), for each of the three models and sample sizes from 75 to 1,000 is shown in Figure 6. The sample variances for each of the BNKDA methods are much larger than those associated with Kruskal's relative importance for $N = 75$, but they are nearly identical for $N = 500$ and sometimes lower when $N = 1,000$.

Figure 6. Sample Variation of Top Driver Identified by BNKDA ($w(l) = l!$) for Varying Sample Sizes Over 1,000 Simulations



BNPLSPM, BNKDA and Kruskal's Relative Importance

With limited simulations for BNPLSPM, it is difficult to draw conclusions about the performance of this methodology. Findings discussed below are only preliminary, but indicate that estimated scores from BNPLSPM align reasonably well with the output of BNKDA and, in terms of top three drivers, are identical to Kruskal's relative importance up to ordering. Similar to Kruskal's relative importance, BNPLSPM tends to reward directly linking nodes with higher driver scores as compared to BNKDA (see X7 in Model 1 or X1 in Model 3 as examples). This does not come as a surprise, as both BNPLSPM and Kruskal's relative importance are regression based methods, while BNKDA relies on conditional independence/dependence.

A comparison of BNKDA and BNPLSPM from Tables 9–11 is summarized below.

- Model 1: Identical top driver, other driver scores are quite different
- Model 2: Identical top three drivers (up to ordering)
- Model 3:
 - Results are quite different when $w(l) = 1$
 - Identical top two drivers with similar score estimates when $w(l) = \sqrt{l}$
 - Identical top three drivers with similar score estimates when $w(l) = l$ or $l!$

Table 9. Model 1 Single Simulation of BNPLSPM Compared to BNKDA and Kruskal's Relative Importance (N = 1,000), Top 3 Drivers Highlighted

Model 1						
Node	BNKDA ($w(l) = l!$)	BNKDA ($w(l) = l$)	BNKDA ($w(l) = \sqrt{l}$)	BNKDA ($w(l) = 1$)	BNPLSPM	Kruskal's relative importance
X1	13.2%	13.2%	16.4%	19.8%	12.9%	7.3%
X2	6.7%	6.7%	8.4%	10.1%	7.6%	2.8%
X3	8.9%	8.9%	11.1%	13.4%	8.9%	2.5%
X4	5.3%	5.3%	6.2%	7.2%	13.0%	13.4%
X5	26.3%	26.3%	23.1%	19.8%	18.4%	22.7%
X6	31.2%	31.2%	27.5%	23.5%	20.3%	31.7%
X7	8.4%	8.4%	7.4%	6.3%	18.9%	19.6%
X8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 10. Model 2 Single Simulation of BNPLSPM Compared to BNKDA and Kruskal's Relative Importance (N = 1,000), Top 3 Drivers Highlighted

Model 2						
Node	BNKDA ($w(l) = l!$)	BNKDA ($w(l) = l$)	BNKDA ($w(l) = \sqrt{l}$)	BNKDA ($w(l) = 1$)	BNPLSPM	Kruskal's relative importance
X1	33.3%	33.3%	41.4%	50.0%	25.2%	11.7%
X2	2.2%	2.2%	1.9%	1.6%	9.5%	8.8%
X3	7.4%	7.4%	6.5%	5.6%	12.6%	11.2%
X4	37.2%	37.2%	32.7%	27.9%	24.7%	38.6%
X5	5.8%	5.8%	5.1%	4.3%	12.0%	11.0%
X6	14.1%	14.1%	12.4%	10.6%	16.0%	18.7%

Table 11. Model 3 Single Simulation of BNPLSPM Compared to BNKDA and Kruskal's Relative Importance (N = 1,000), Top 3 Drivers Highlighted

Model 3						
Node	BNKDA ($w(l) = l!$)	BNKDA ($w(l) = l$)	BNKDA ($w(l) = \sqrt{l}$)	BNKDA ($w(l) = 1$)	BNPLSPM	Kruskal's relative importance
X1	0.2%	0.2%	0.1%	0.1%	5.5%	3.6%
X2	12.6%	11.6%	13.3%	14.3%	7.1%	7.2%
X3	3.6%	6.6%	9.2%	12.3%	7.5%	2.3%
X4	0.6%	3.5%	5.6%	8.6%	2.6%	0.4%
X5	25.3%	23.1%	18.7%	14.3%	18.4%	33.0%
X6	10.8%	9.9%	11.3%	12.3%	11.9%	6.9%
X7	2.5%	4.6%	6.5%	8.6%	5.0%	1.3%
X8	21.6%	19.8%	16.0%	12.3%	17.0%	23.1%
X9	7.6%	6.9%	7.9%	8.6%	10.5%	6.7%
X10	15.2%	13.9%	11.2%	8.6%	14.5%	15.5%

DISCUSSION AND SUMMARY

Differences from Traditional KDA Methodologies

While BNKDA will often identify the same top drivers as more common techniques, there are several differences between the methodologies.

- When using BNs with KDA, scores estimated from BNKDA align with the BN, telling a single, cohesive story.
- BNKDA makes use of the interrelationships between drivers when estimating scores, while traditional techniques assume only a direct relationship from driver to target.
- BNKDA is able to handle multiple target attributes simultaneously, while other methods would require separate models.
- The resulting estimates from traditional techniques, when compared to those of BNKDA, tend to favor attributes with a direct relationship to the target.
- BNKDA uses a series of conditional independence tests and a greedy hill-climbing search algorithm based on the BIC, while traditional methods rely on regression theory.
- The alternative fitting technique specific to BNKDA may yield estimates when other models fail due to multicollinearity or other strict assumptions (but use caution in these cases as the estimates may be unstable).
- The MMHC algorithm to fit a BN is prone to completely removing some attributes from the model, resulting in driver estimates of 0.

Weight Factor Recommendation

Evidence from the simulation study indicates that a weight factor is necessary to avoid ambiguity in driver estimates and counterintuitive results. Findings from the weight comparison point to the choice of weight factor $w(l) = l!$ for the following reasons:

- Provides adequate differentiation among drivers
- Driver estimates are most consistent with the other methods
- Emphasizes attributes with shorter paths to the target, i.e., attributes that will impact the target in a more direct fashion

Future Exploration

While the simulation study shows promising results, there is still a lot to explore with BNKDA.

- There is a wide array of methods available to calculate arc strength. What impact would different methodologies have on the overall driver estimates?
- Changing the arc strength standardization procedure would change the driver estimates, but how?
- How would a different path strength aggregation function other than the sum (such as the mean or median) impact driver estimates?

These are some open areas reserved for future research.

Summary

Traditional KDA methods present a single driver score for each attribute, providing probable candidates for key drivers, but offer little to no information about the interrelationships between them. Bayesian networks provide insights into those relationships via a DAG, but provide limited guidance on a set of key features. The proposed algorithm, BNKDA, links the methodologies together by calculating driver estimates directly from a fitted BN. This produces both the network visualization and driver scores to pinpoint key attributes.

While further study is required to fully understand the behavior of BNKDA, results from the simulation study indicate that implementation of this technique will provide similar results to traditional KDA methods, without emphasizing attributes that directly influence the target. The additional insights gained from BNKDA make this a useful addition to the KDA toolbox.



Benjamin Cortese

APPENDIX

All models are simulated using the statistical language R. Each hierarchical model is defined using the notation $N(\mu, \sigma)$ for a normal distribution with mean μ and variance σ and $\text{Binom}(n, p)$ for a binomial distribution with n trials and success probability p . The intended arc structure for each model is provided in Tables 12–14.

Model 1

Table 12. Intended Arc Structure for Model 1

Model 1 Arc Structure	
From	To
X5	Y
X6	Y
X7	Y
X4	Y
X4	X7
X1	X5
X2	X6
X3	X6

$$\begin{aligned}
 x_1 &\sim N(6, 1) \\
 x_2 &\sim N(7, 1) \\
 x_3 &\sim N(8, 1) \\
 x_4 &\sim N(2, 1) \\
 x_8 &\sim N(5, 1) \\
 x_5 &\sim N(x_1, 1) \\
 x_6 &\sim N\left(\frac{(x_2 + x_3)}{2}, 1\right) \\
 x_7 &\sim N(x_4, 1) \\
 y &= -2 + (x_4 + 1.25x_5 + 1.5x_6 + x_7)/4 + \epsilon \\
 \epsilon &\sim N(0, 1)
 \end{aligned}$$

After each of the above are generated from the specified normal distribution, the values are rounded to the nearest integer and coerced to the interval $[1, 10]$ to simulate a 10-point scale survey response.

Model 2

Table 13. Intended Arc Structure for Model 2

Model 2 Arc Structure	
From	To
X1	X2
X1	X3
X1	X4
X1	X5
X1	X6
X2	Y
X3	Y
X4	Y
X5	Y
X6	Y

$$x_1 \sim \text{Binom}(9, 0.6) + 1$$

$$x_2 \sim \begin{cases} x_1 \leq 3, \text{Binom}(9, 0.4) + 1 \\ 3 < x_1 \leq 6, \text{Binom}(9, 0.7) + 1 \\ x_1 > 6, \text{Binom}(9, 0.9) + 1 \end{cases}$$

$$x_3 \sim \begin{cases} x_1 \leq 5, \text{Binom}(9, 0.6) + 1 \\ 5 < x_1 \leq 8, \text{Binom}(9, 0.75) + 1 \\ x_1 > 8, \text{Binom}(9, 0.8) + 1 \end{cases}$$

$$x_4 \sim \begin{cases} x_1 \leq 4, \text{Binom}(9, 0.5) + 1 \\ 4 < x_1 \leq 8, \text{Binom}(9, 0.6) + 1 \\ x_1 > 8, \text{Binom}(9, 0.9) + 1 \end{cases}$$

$$x_5 \sim \begin{cases} x_1 \leq 2, \text{Binom}(9, 0.3) + 1 \\ 2 < x_1 \leq 6, \text{Binom}(9, 0.5) + 1 \\ x_1 > 6, \text{Binom}(9, 0.7) + 1 \end{cases}$$

$$x_6 \sim \begin{cases} x_1 \leq 3, \text{Binom}(9, 0.5) + 1 \\ 3 < x_1 \leq 8, \text{Binom}(9, 0.65) + 1 \\ x_1 > 8, \text{Binom}(9, 0.8) + 1 \end{cases}$$

$$y \sim \text{Binom}(9, p) + 1 \text{ where } p = \prod_{i=2}^6 p_i \text{ and}$$

$$p_2 = \begin{cases} x_2 \leq 6, 0.85 \\ 6 < x_2 \leq 9, 0.9 \\ x_2 > 9, 1 \end{cases}$$

$$p_3 = \begin{cases} x_3 \leq 6, 0.8 \\ 6 < x_3 \leq 8, 0.85 \\ x_3 > 8, 0.9 \end{cases}$$

$$p_4 = \begin{cases} x_4 \leq 5, 0.75 \\ 5 < x_4 \leq 7, 0.8 \\ x_4 > 7, 0.95 \end{cases}$$

$$p_5 = \begin{cases} x_5 \leq 6, 0.9 \\ 6 < x_5 \leq 8, 0.95 \\ x_5 > 8, 1 \end{cases}$$

$$p_6 = \begin{cases} x_6 \leq 5, 0.85 \\ 5 < x_6 \leq 7, 0.95 \\ x_6 > 7, 1 \end{cases}$$

Model 3

Table 14. Intended Arc Structure for Model 3

Model 3 Arc Structure	
From	To
X1	Y
X5	Y
X8	Y
X10	Y
X2	X5
X6	X8
X9	X10
X3	X6
X7	X9
X4	X7

$$x_1 \sim \text{Binom}(9, 0.1) + 1$$

$$x_2 \sim \text{Binom}(9, 0.3) + 1$$

$$x_3 \sim \text{Binom}(9, 0.5) + 1$$

$$x_4 \sim \text{Binom}(9, 0.7) + 1$$

$$x_5 \sim \begin{cases} x_2 \leq 3, \text{Binom}(9, 0.5) + 1 \\ 3 < x_2 \leq 5, \text{Binom}(9, 0.6) + 1 \\ x_2 > 5, \text{Binom}(9, 0.75) + 1 \end{cases}$$

$$x_6 \sim \begin{cases} x_3 \leq 4, \text{Binom}(9, 0.4) + 1 \\ 4 < x_3 \leq 7, \text{Binom}(9, 0.7) + 1 \\ x_3 > 7, \text{Binom}(9, 0.9) + 1 \end{cases}$$

$$x_7 \sim \begin{cases} x_4 \leq 7, \text{Binom}(9, 0.6) + 1 \\ 7 < x_4 \leq 8, \text{Binom}(9, 0.7) + 1 \\ x_4 > 8, \text{Binom}(9, 0.85) + 1 \end{cases}$$

$$x_8 \sim \begin{cases} x_6 \leq 5, \text{Binom}(9, 0.45) + 1 \\ 5 < x_6 \leq 7, \text{Binom}(9, 0.65) + 1 \\ x_6 > 7, \text{Binom}(9, 0.85) + 1 \end{cases}$$

$$x_9 \sim \begin{cases} x_7 \leq 6, \text{Binom}(9, 0.75) + 1 \\ 6 < x_7 \leq 8, \text{Binom}(9, 0.85) + 1 \\ x_7 > 8, \text{Binom}(9, 0.95) + 1 \end{cases}$$

$$x_{10} \sim \begin{cases} x_9 \leq 7, \text{Binom}(9, 0.6) + 1 \\ 7 < x_9 \leq 8, \text{Binom}(9, 0.8) + 1 \\ x_9 > 8, \text{Binom}(9, 0.95) + 1 \end{cases}$$

$$y \sim \text{Binom}(9, p) + 1 \text{ where}$$

$$p = \prod_{i=1,5,8,10} p_i \text{ and}$$

$$p_1 = \begin{cases} x_1 \leq 2, 0.8 \\ 2 < x_1 \leq 3, 0.85 \\ x_1 > 3, 0.95 \end{cases}$$

$$p_5 = \begin{cases} x_5 \leq 5, 0.8 \\ 5 < x_5 \leq 7, 0.9 \\ x_5 > 7, 1 \end{cases}$$

$$p_8 = \begin{cases} x_8 \leq 5, 0.75 \\ 5 < x_8 \leq 8, 0.9 \\ x_8 > 8, 0.95 \end{cases}$$

$$p_{10} = \begin{cases} x_{10} \leq 8, 0.85 \\ 8 < x_{10} \leq 9, 0.95 \\ x_{10} > 9, 1 \end{cases}$$

REFERENCES

- Daly, R., & Shen, Q. (2007). Methods to Accelerate the Learning of Bayesian Network Structures. Proceedings of the 2007 UK Workshop on Computational Intelligence. London: Imperial College.
- Ebert-Uphoff, I. (2007). Measuring Connection Strengths and Link Strengths in Discrete Bayesian Networks. Georgia Tech Research Report.
- Grömping, U. (2006, 09 11). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software*, 17(1), 1–27.
- Kruskal, W. (1987). Relative Importance by Averaging over Orderings. *The American Statistician*, 41(1), 6–10.
- Margaritis, D. (2003, May). Learning Bayesian Network Model Structure from Data. Pittsburgh, PA, USA: School of Computer Science, Carnegie-Mellon University.
- Nicholson, A.E. (1998). Using Mutual Information to determine Relevance in Bayesian Networks. In H.-Y.L. Motoda, *PRICAI'98: Topics in Artificial Intelligence* (Vol. 1531, pp. 399–410). Berlin: Springer Berlin Heidelberg.
- Ringle, C.M., Wende, S., & Becker, J.-M. (2015). SmartPLS 3. SmartPLS GmbH. Retrieved from <http://www.smartpls.com>
- Russell, S.J. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall.
- Scutari, M. (2010). Learning Bayesian Networks with bnlearn R Package. *Journal of Statistical Software*, 35(3), 1–22. <http://www.jstatsoft.org/v35/i03/>
- Tsamardinos, I., Aliferis, C., & Statnikov, A. (2003). Algorithms for Large Scale Markov Blanket Discovery. Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference (pp. 376–380). St. Augustine: AAAI Press.
- Tsamardinos, I., Brown, L., & Aliferis, C. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1), 31–78.
- Vinzi, V., Trinchera, L., & Amato, S. (2010). PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement. In V. Vinzi, W. Chin, J. Henseler, & H. Wang, *Handbook of Partial Least Squares: Concepts, Methods and Application* (pp. 47–82). Berlin Heidelberg: Springer.
- Zuk, O., Margel, S., & Domany, E. (2012, June 27). On the Number of Samples Needed to Learn the Correct Structure of a Bayesian Network. Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence.

PRODUCT RELEVANCE AND NON-COMPENSATORY CHOICE

MARC R. DOTSON

BRIGHAM YOUNG UNIVERSITY

ROGER A. BAILEY

GREG M. ALLENBY

THE OHIO STATE UNIVERSITY

1 INTRODUCTION

Consumers enter the marketplace to find products that will serve as solutions in helping them address their needs or accomplish their goals. Products are composed of a variety of features or attributes. A consumer uses these attributes to infer the effectiveness of a given product to serve as a solution. A product that a consumer believes will be able to help address their specific needs and goals is *relevant* to that consumer. However, not all attributes are used in the same way to determine product relevance. Furthermore, the way consumers identify product relevance reveals information about the needs they want to address or the goals they seek to accomplish.

Most of the attributes that compose a given product serve tangible functions (e.g., flavor). At least two of the attributes are different: price and brand. Price can serve as a signal of quality but primarily determines how a product fits within a consumer's budget and the greater basket of products. Brand, out of all attributes, is uniquely tied to consumer inference about product relevance. The consumer enters a marketplace with beliefs about brands already formed, including each brand's effectiveness in serving as a solution for the consumer's needs and goals. Thus, brand beliefs can serve as a simple heuristic in order to shortcut consumers' decision-making process and reduce their cognitive load. In other words, brand alone can be used to determine product relevance.

Attributes beyond brand and price can also be used to determine product relevance. In particular, the presence of certain attribute levels may lead a consumer to perceive a given product as able to address their specific needs and goals despite what they believe about the product's brand. In this way, the determination of product relevance is a sub-compensatory process. Either a product's brand is enough for a consumer to infer product relevance or the presence of certain attribute levels leads a consumer to infer product relevance. We develop a model that allows us to capture these two ways to product relevance as part of an extended model of choice.

The heuristics consumers adopt in order to shortcut the decision-making process and reduce cognitive load are an essential feature of screening models in the non-compensatory choice literature (Aribarg, Otter, Zantedeschi, et al. 2018). We adapt the structure of these models as a way to uncover the particular attribute levels that lead to product relevance beyond pre-existing brand beliefs. In order to employ a screening model in this way, we make two critical assumptions. First, consumers consider and choose relevant products. Second, attribute levels that are *not* used to screen (i.e., remove products from consideration) are those that help a consumer infer product relevance.

Firms benefit from understanding which attributes drive product relevance, and thus consideration. This is especially true for brands without strong existing loyalties and when those

attributes that drive product relevance are different from those that lead to choice. Furthermore, understanding which attributes drive product relevance reveals something about consumers' needs and goals that drive them to the marketplace to begin with. Answers with respect to the *why* of product relevance can inform product development and promotion strategies in a way to engender brand loyalty and inform consumers' brand beliefs.

The remainder of the paper will be organized as follows. In Section 2, we consider model-free, empirical evidence with respect to this idea of product relevance and its place in consumer choice. In Section 3, we specify our model. In Section 4, we provide results. In Section 5, we conclude.

2 EMPIRICAL EVIDENCE

Consumers determine product relevance based either on existing brand beliefs or the presence of certain attribute levels. Before specifying a model to untangle these two ways to infer product relevance, we consider empirical evidence that illustrates the importance of its development. In particular, with information on brand beliefs, we can consider model-free evidence with respect to choices made for products that are or are not “brand relevant” (i.e., relevant based on brand beliefs).

We collected conjoint data in the premium chocolate category. Our 788 respondents completed 10 choice tasks, each with four product alternatives plus an outside option. We specified nine attributes, including brand and price, with a total of 70 attribute levels. Prior to the conjoint, respondents were asked to indicate which chocolate brands they would consider purchasing.

Using this stated brand consideration (i.e., brand belief) information, we can count how many of the chosen alternatives were brand relevant. Figure 1 shows that less than half of the chosen alternatives are brand relevant. If brand beliefs were strong enough, we might expect to see all of the chosen alternatives as brand relevant, particularly because respondents could always pick the outside option (i.e., the “none” option) if a brand-relevant alternative wasn't included in a given choice task. However, that's not what we observe in aggregate.

The proportion of brand-relevant chosen alternatives for each respondent is shown in Figure 2. This is a tightly-packed bar plot where each respondent has their own bar, sorted by the proportion of brand-relevant chosen alternatives. We can see there is a subset of respondents for whom all of their chosen alternatives are brand relevant (the section of respondents with complete dark bars on the left) as well as a subset of respondents for whom none of their chosen alternatives are brand relevant (the section of respondents with complete light grey bars on the right). For this second group, the subset of respondents for whom none of their chosen alternatives are brand relevant, 86% of them did say prior to the conjoint that at least one brand was relevant, but still never selected a brand-relevant alternative.

Figure 1

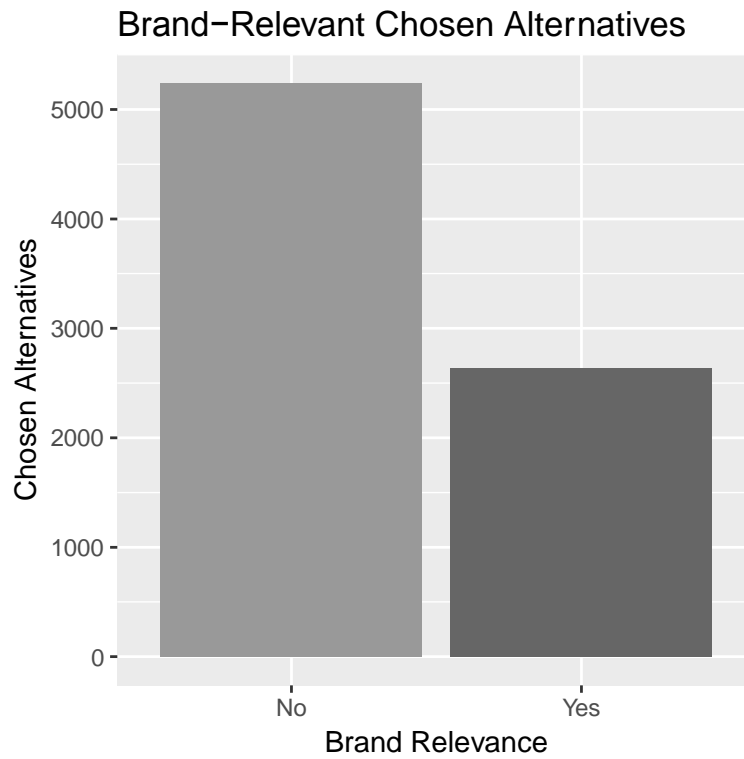
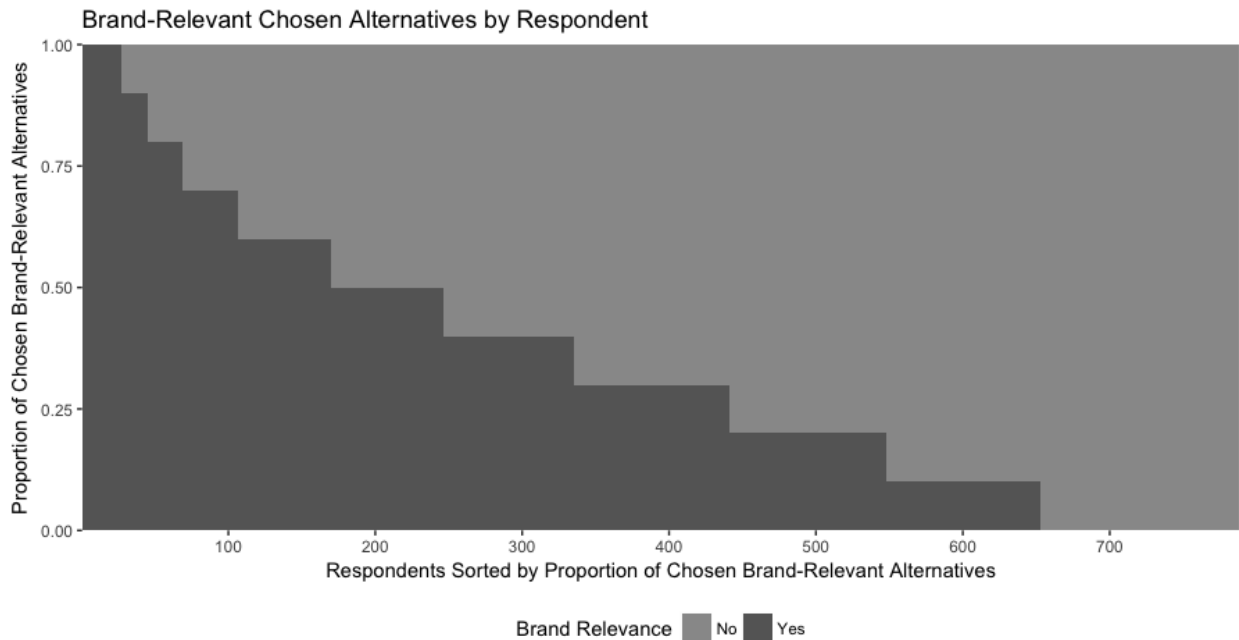


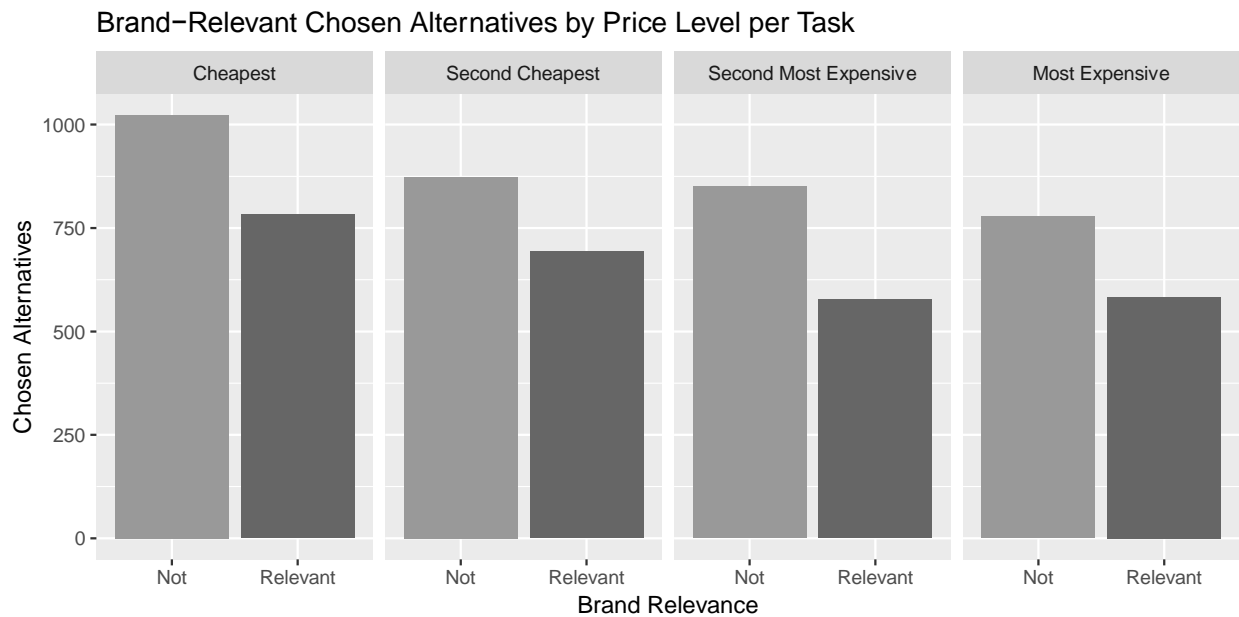
Figure 2



We see that respondents are choosing alternatives with brands they initially said they would not consider. One final model-free check would be to determine whether or not this behavior is simply driven by price. In other words, when presented with a more expensive, brand-relevant

alternative and a less expensive, brand-irrelevant alternative, are respondents more likely to choose the latter? In Figure 3, we see the number of chosen alternatives that are and are not brand relevant split by the *relative* price level for each choice task. We can clearly see that the more relatively expensive alternatives are chosen less frequently. However, if price was driving respondents to pick brand-irrelevant alternatives, we would see a disproportionate increase in the number of brand-irrelevant choices as the price gets cheaper. We do not observe this.

Figure 3



Something is driving respondents to choose things that are not brand relevant, but it isn't price. The only other explanation is that there are certain attribute levels present that lead respondents to see product alternatives as relevant, despite their belief about the associated brand. With this empirical evidence in hand, we will now walk through the development of a model to disentangle the two ways to product relevance.

3 MODEL SPECIFICATION

Non-compensatory choice models provide a more realistic description of consumer choice. Such models specify a two-stage decision process that begins with consideration and ends with choice. In these models, consumers typically screen products (i.e., remove them from consideration) based on certain attribute levels, which are identified using chosen products alone. The screening mechanism is assumed to be the same for each attribute. Finally, the propensity to screen based on a given attribute level is typically homogeneous. We develop a non-compensatory choice model that identifies screening based on chosen products and brand beliefs, includes two different ways for products to be considered, and allows for the propensity to screen to be heterogeneous.

Our interest is which attribute levels are being used to infer product relevance and which products are being considered. As previously stated, we assume that products that are relevant to the consumer are considered and chosen. We also assume that attribute levels that are *not* being used to screen (i.e., remove products from consideration) help consumers infer product

relevance. We adapt the structure of a non-compensatory screening model in order to uncover those attributes that drive product relevance outside of brand beliefs.

Conjunctive Screen

The standard random utility model is fully compensatory. No matter how much a consumer dislikes a certain attribute level, this disutility can be compensated for by the presence of attribute levels that they highly prefer. The standard non-compensatory model adds a simple constraint such that if certain attribute levels are present for a product alternative for a given respondent, they will never choose that alternative. In other words, the inclusion of their most-preferred attribute levels will do nothing to compensate for the presence of the attribute levels they are screening on. The simplest way this non-compensatory or screening process is represented is by what is called a conjunctive screen: an alternative is considered as long as *none* of the attribute levels a given respondent is screening on are present.

This conjunctive screening model adds a set of respondent-level screening parameters to the model beyond the standard respondent-level part-worth utilities. These screening parameters are binary, with a separate parameter associated with each of the attribute levels. A screening parameter of one indicates the respondent screens on that attribute level. Thus, if that attribute level (or any other attribute level the respondent may be screening on) is present in an alternative, that respondent will not choose it—the product is removed from consideration.

Conjunctive screens are identified by the attribute levels in the chosen or picked alternatives. If an attribute level is ever in the picked alternatives, we know it isn't being used to screen; if it's never in the picked alternatives, it may or may not be used to screen and its probability of being used to screen is drawn from a Bernoulli distribution with a parameter representing a homogeneous propensity to screen. This is depicted in Figure 4, where $\tau_{h,l}$ is the screening parameter for respondent h and attribute level l and θ_l is the propensity to screen on attribute level l across respondents.

Figure 4

<i>Attribute Level</i>	
Picked	$\tau_{h,l} = 0$
Never Picked	$\tau_{h,l} \sim \text{Bernoulli}(\theta_l)$

Disjunction of Conjunctive Screens

Our proposed non-compensatory choice model uses a *disjunction of conjunctive* screens where an alternative is considered when none of the attribute levels a given respondent is screening on are present *or* the given respondent is explicitly considering the brand. It is this *or* that creates a disjunction. These two ways to screen products out of consideration describe the two ways product relevance is inferred. A product is relevant (i.e., not screened from consideration) if the given respondent believes the brand can serve as a solution *or* the presence of certain attribute levels (i.e., ones with screening parameters of zero) lead the given respondent to believe the product can serve as a solution, regardless of their brand beliefs. The identification of the screening parameters in this proposed model is slightly more complicated given the disjunction, as depicted in Figure 5.

Figure 5

<i>Attribute Level</i>	<i>Brand Relevant</i>	<i>Not Brand Relevant</i>
Picked	$\tau_{h,l} \sim \text{Bernoulli}(\theta_l)$	$\tau_{h,l} = 0$
Never Picked	$\tau_{h,l} \sim \text{Bernoulli}(\theta_l)$	$\tau_{h,l} \sim \text{Bernoulli}(\theta_l)$

Additionally, we would like to allow for the propensity to screen based on a given attribute level (i.e., θ_l) to be heterogeneous (i.e., $\theta_{h,l}$). This is accomplished by including an upper-level model for each of the propensities to screen variables so that information can be shared across respondents, with a function similar to the usual HB upper-level model over the part-worth utilities.

4 RESULTS

Again, our conjoint data is from the premium chocolate category. Our 788 respondents completed 10 choice tasks, each with four product alternatives plus an outside option. We specified nine attributes, including brand and price, with a total of 70 attribute levels, a subset of which are provided in Figure 6. Prior to the conjoint, respondents were asked to indicate which chocolate brands they would consider purchasing, giving us information to drive the brand portion of the proposed disjunction of conjunctive screens.

Figure 6

<i>Attributes (# Levels)</i>	<i>Levels</i>			
Brand (10)	Lindt	Godiva	Ghiradelli	...
Shape (6)	Cover	Round	Crown	...
Filling Description (8)	Cream	Mousse	Ganache	...
Chocolate Flavor (4)	Dark Chocolate	Milk Chocolate	White Chocolate	...
Filling Flavor (12)	Chocolate	Vanilla	Caramel	...
Price Per Ounce (11)	\$0.50	\$0.60	\$0.70	...
Pieces (5)	1 pc.	4 pcs.	10 pcs.	...
Packaging Type (9)	Resealable Bag	Bag	Flat Box	...
Wrapping Type (5)	Flow Packed	Single Twist	Double Twist	...

We ran five different models. The HMNL model is the standard, fully compensatory choice model. The Conjunctive Screen with θ_l is the standard screening model with homogeneous propensity to screen on each of the attribute levels. The Conjunctive Screen with $\theta_{h,l}$ is a screening model modified to allow for heterogeneous propensity to screen on each of the attribute levels. The Disjunction of Conjunctive Screens with θ_l is a version of our proposed model simplified by assuming a homogeneous propensity to screen. The Disjunction of Conjunctive Screens with $\theta_{h,l}$ is our complete proposed model. Model results are provided in Figure 7.

Figure 7

<i>Model</i>	In-Sample		Out-of-Sample	
	<i>LMD</i>	<i>MSE</i>	<i>Hit Rate</i>	<i>Hit Prob.</i>
HMNL	-3472.48	3.168	0.333	0.327
Conjunctive Screen with θ_l	-2200.66	3.125	0.306	0.303
Conjunctive Screen with $\theta_{h,l}$	-2605.25	3.111	0.311	0.306
Disjunction of Conjunctive Screens with θ_l	-2986.97	3.056	0.338	0.332
Disjunction of Conjunctive Screens with $\theta_{h,l}$	-3264.69	3.121	0.324	0.319

In-sample fit is calculated using the log-marginal density (closer to zero is better). Out-of-sample fit on a single, randomly selected choice task for each respondent is calculated using mean squared error, hit rate, and hit probability. While the standard screening model, which is easiest to identify, performs best in terms of in-sample fit, the simplified version of our proposed model performs best in terms of out-of-sample fit. This evidence suggests that these two ways to infer product relevance are indeed present in our data. Furthermore, the reason the complete proposed model doesn't perform better is likely out of the difficulty of identifying respondent-level propensity to screen parameters given the shallowness of the data (i.e., only 10 choice tasks per respondent over 70 attribute levels).

More than building a better choice model, the proposed non-compensatory choice model also allows us to consider the drivers of product relevance vs. the drivers of choice. The estimates of the propensity to screen variables suggest that Filling Flavor (Caramel, Peach, Hazelnut, and Peanut Butter) is the most important driver of product relevance via attributes. In contrast, the part-worth utility estimates suggest that Brand, Price, and Packaging Type are the most important drivers of choice. That Filling Flavor leads to product relevance suggests the needs or goals consumers have with respect to the premium chocolate category is focused on hedonic consumption. This isn't surprising, but it's something that we may have downplayed if all we considered were the part-worth utilities. If a firm is trying to move into this category with a brand that lacks loyalty, clearly the right Filling Flavor can help it be initially considered by consumers.

5 CONCLUSION

Consumers consider and thus purchase products that are relevant to them—either because of their brand beliefs or the presence of certain attribute levels. Separating and uncovering the drivers of product relevance allow firms to understand something of the underlying motivations driving consumers into the marketplace to begin with. This knowledge will help firms to design promotions and products that address those motivations, build brand loyalty, and inform consumers' brand beliefs.



Marc R. Dotson



Roger A. Bailey



Greg M. Allenby

REFERENCES

Anocha Aribarg, Thomas Otter, Daniel Zantedeschi, Greg Allenby, Taylor Bentley, David Curry, Marc Dotson, Ty Henderson, Elisabeth Honka, Rajeev Kohli, Kamel Jedidi, Stephan Seiler, and Xin Wang (2018), “Advancing Non-Compensatory Choice Models in Marketing,” *Customer Needs and Solutions*, 5(1–2), 82–92.