PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

March 2012

FOREWORD

These proceedings are a written report of the sixteenth Sawtooth Software Conference, held in Orlando, Florida, March 21-23, 2012. Two-hundred fifty attendees participated, representing record attendance for this event. This conference was held in concert with the fourth Conjoint Analysis in Healthcare Conference, chaired by John Bridges of Johns Hopkins University. Conference sessions were held at the same venue and ran concurrently. (These proceedings contain only the papers delivered at the Sawtooth Software Conference.)

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, menu-based choice, cluster ensemble analysis, design of choice experiments, Genetic Algorithm searches, and text analytics.

The papers and discussant comments are in the words of the authors and very little copy editing was performed. We are grateful to these authors for continuing to make this conference a valuable event and advancing our collective knowledge in this exciting field.

Sawtooth Software

September, 2012

CONTENTS

SUMMARY OF FINDINGS

The sixteenth Sawtooth Software Conference was held in Orlando, Florida, March 21-23, 2012. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2012 Sawtooth Software Conference Proceedings.

Game Theory and Conjoint Analysis: Using Choice Data for Strategic Decisions (Christopher Chapman, Google, and Edwin Love, Western Washington University): Christopher and Edwin encouraged the audience to think about using game theory to leverage the benefits of conjoint analysis with management, rather than just presenting variations from base case simulation results, or especially conjoint utilities and importances. With game theory, the researcher and management consider the possible changes to the client's product positioning and/or price, and the likelihood of possible competitive responses. The expected value of each outcome (in terms of share, revenues, or profits) can be simulated using the market simulator. Focusing on game theory within a conjoint analysis encourages the entire team to agree on specific goals and outcomes. By using game theory, the researcher isn't just presenting research findings to management, but is participating in developing effective business strategy. From a practical point of view, Christopher and Edwin emphasized that it's important to get good quality conjoint measurement, to include significant interactions in the model between brand and price if warranted, and obtain a better gauge of the "None" response than the traditional None within CBC.

Contrast Effects on Willingness to Pay: How Conjoint Design Affects Adult Day Care Preferences (David Bakken, Michaela Gascon, and Dan Wasserman, KJT Group): Many market research studies seek to quantify respondents' willingness to pay (WTP) for a product or specific feature. Unfortunately, we must simplify reality in order to estimate WTP. David explained that seemingly small contextual clues we present to respondents can have a large effect on the estimated WTP. He and his co-authors conducted an experiment using the van Westendorp pricing meter technique (which they felt was particularly susceptible to context effects), and found that providing respondents with different reference price points, and also an "outside good" comparison, can strongly affect the WTP outcome. In a second study, CBC was also shown to be affected by contextual clues, including pre-conditioning on attributes, range of levels included for price, and the specification of the "outside good" (the None category). David concluded by emphasizing how important it is to make a CBC study as realistic as possible, with realistic reference information and level ranges. Comparisons to "outside goods" should be explicit.

Optimizing Pricing of Mobile Apps with Multiple Thresholds in Anchored MaxDiff (Paul Johnson, Survey Sampling International and Brent Fuller, The Modellers): Paul showed results of an internal Survey Sampling International (SSI) study that sought to measure the interest among SSI customers in buying a mobile SSI app. Paul used anchored MaxDiff, where both indirect (Louviere method) and direct (Lattery) methods were used to rescale MaxDiff utilities with respect to a two buy/don't buy thresholds (one buy for free, and one buy for $1). Furthermore, Paul and Brent experimentally manipulated whether the mobile app was shown first as free or as $1.00. Anchor points, Paul explained, are really just additional items that are included in the MaxDiff design. Multiple price points could constitute these thresholds, or even anchoring descriptors such as "consider" vs. "purchase" or "important" vs. "essential". Paul and

Brent found quite large differences in the projected demand for the mobile app depending on whether respondents completed the indirect or direct elicitation methods for the buy/no buy threshold. The direct method resulted in much lower (and in the authors' opinion more realistic) projected take rates than the indirect method. The indirect method was particularly sensitive to the order in which the prices for the app were presented to respondents.

Continued Investigation into the Role of the "Anchor" in MaxDiff and Related Tradeoff Exercises (Jack Horne, Bob Rayner, Reg Baker, and Silvo Lenart, Market Strategies International): Jack reiterated that the main goal of anchored MaxDiff is to remove the limitation of the relative scaling of items in traditional MaxDiff to be able to assess, for example, whether a respondent thinks that most of the items are good/important or most of them are bad/unimportant. They tested multiple ways of estimating the threshold, including direct and indirect methods, MaxDiff vs. paired comparisons, and another technique his firm calls two-by-four. A split-sample experiment was set up with real respondents, and the results were assessed in terms of different measures of success. The authors also investigated different ways to code the anchoring information prior to HB estimation. The indirect and direct methods produced quite different absolute positioning of the threshold parameter, with the indirect method showing that many more items exceed the threshold (many more items are "important"). The direct method also resulted in quicker interviews with fewer respondents dropping out. The authors also found the pairwise comparisons to perform admirably compared to MaxDiff. In conclusion, the authors cautioned that while there are benefits to anchored MaxDiff (or anchored paired comparisons), the issue of scale use bias reenters the equation. Specifically, they demonstrated that the utility of the threshold differs strongly by country. So, cross-cultural comparisons (a strength of traditional MaxDiff) are hampered by using anchored scaling.

Using MaxDiff for Evaluating Very Large Sets of Items (Ralph Wirth and Anette Wolfrath, GfK Marketing Sciences): While MaxDiff has become a very popular method in our community, the challenge of how to deal with very large lists of items (such as more than 60) is often raised. Should researchers just include all the items in MaxDiff, even though the data matrix becomes very sparse at the individual level? Can HB be used effectively in the case of such sparse individual-level designs? Ralph presented results demonstrating that HB can do a creditable job of estimation, even given data conditions that are thinner than Sawtooth Software has recommended within its MaxDiff software documentation. The authors used both simulated data and empirical results from a split-sample design with real respondents. They compared designs which showed only a subset of the items to each respondent vs. those that showed all items (fewer times) to each respondent. Their results suggested that showing all items fewer times to each respondent was probably a little bit better than showing a subset of the items to each respondent, where each item appeared more times. But, critically, the Degrees of Freedom advanced setting in CBC/HB software had a large effect on the ability to obtain quality results in the face of sparse designs. Specifically, the D.F. setting needed to be increased. An augmented MaxDiff approach (using an initial Q-Sort phase) was shown to perform extremely well at the individual level, but it has the drawback of being more complicated to design and program.

What's in a Label? Business Value of "Soft" vs "Hard" Cluster Ensembles (Nicole Huyghe and Anita Prinzie, solutions-2): Ensemble Analysis is a relatively new technique for the marketing research community, and is a more sophisticated way to do clustering analysis. Inputs to ensemble analysis have often been discrete membership in a candidate set of segmentation solutions. Nicole and Anita investigated what would happen if probabilistic membership (soft cluster ensembles) rather than discrete memberships (hard ensembles) were used in the ensemble. With a soft candidate segmentation solution, a respondent may be 70% likely to be in group 1, 20% in group 2, and 10% in group 3, etc. Soft memberships contain more statistical information, so it was hypothesized that they might perform better. Using three industry data sets, they compared hard to soft clustering in terms of stability, cluster integrity, accuracy, and size. They found that hard clustering tended to yield more heterogeneous cluster solutions, and more balanced segments (lower tendency to result in very small or very large segments). But, when the ensemble included a greater number of similar solutions, soft ensembles were more stable. There wasn't enough evidence from their research to conclude that one approach for ensemble analysis was clearly superior to the other.

How Low Can You Go? Toward a Better Understanding of the Number of Choice Tasks Required for Reliable Input to Market Segmentation (Jane Tang and Andrew Grenville, Vision Critical): A few papers presented at past Sawtooth Software conferences have suggested that researchers can obtain solid results for CBC studies using 10 or fewer choice tasks per respondent. But, these earlier studies did not consider how many tasks are needed to reliably assign respondents within a market segmentation. Jane presented results from seven industry studies, where the ability to re-assign respondents reliability into clusters developed given full information was tested, using a reduced number of choice tasks per respondent. Her findings suggest that the clustering process often is unstable even with large numbers of choice tasks. And, segmenting respondents into a larger number of clusters requires more choice tasks than when assigning respondents to fewer number of clusters. She concluded that under most conditions, segmentation requires approximately 10 choice tasks per respondent in CBC. Although she used HB utilities and Cluster Analysis to support her findings, Jane does not suggest that HB+Cluster is superior to directly segmenting via Latent Class Her paper simply demonstrates that reducing the number choice tasks from 16 to 10 does not seem to hinder the process of creating market segmentation.

"The Individual Choice Task Threshold" Need for Variable Number of Choice Tasks (Peter Kurz, TNS Infratest Forschung GmbH, and Stefan Binner, bms - marketing research + strategy): Peter reflected on the fact that CBC is the most widely used conjoint methodology in our industry, and that in the authors' experience, fewer tasks are often all that is needed (fewer than typical norms in the industry). Furthermore, with the wide-spread use of panel sample, researchers, panel vendors, and respondents are pushing for fewer tasks in CBC questionnaires. And, based on their experience not only with the data, but by observing individual respondents complete CBC surveys, they believe that many respondents reach a point in the CBC survey where they disengage and begin to give data of questionable quality. They re-analyzed twelve commercial CBC data sets to see if they could use quantitative measures to somehow detect the threshold point at which respondents disengaged. Their idea is that if somehow this could be detected in real-time, then the survey could stop asking more tasks, respondent burden could be reduced, and fewer overall choice tasks would need to be asked for the sample, while maintaining equal or better results. Much of their analysis focused on measures of internal fit to the tasks used in estimation (RLH), which audience members pointed out may not be the best approach. Using the existing data sets, the authors demonstrated that strategically throwing away 38% of the choice tasks would not lead to a large decrement in predictive quality of the CBC models. But, how precisely to do this in real time and apply it during data collection is still a problem to be solved, with no clear answers at this point. A possibility is to discard the less-reliable later tasks after the fact, when the threshold for individual respondents has been identified during the analysis.

Taking Nothing Seriously or "Much Ado About Nothing" (Kevin Karty and Bin Yu, Affinnova): Kevin reviewed the common ways the None alternative is used within CBC questionnaires: as an available concept, as multiple available None concepts (each with a different label), as dual-response 2-point scale (would buy/would not buy), and also as a dual-response 5-point scale (definitely, probably, might or might not, probably would not, and definitely would not buy). The standard None approach results in the lowest None usage by respondents. The use of None increases in later tasks. The dual-response None leads to significantly higher None usage. Still, Kevin argued, the 2-point dual-response None leads to too much exaggeration of purchase intent by the respondents. Kevin suggested that the 5-point dual-response scale, where top-box only indicates a "buy" (and bottom 4 boxes indicate "None") leads to more realistic measures of actual purchase intent for the sample. Kevin also reasoned that the None alternative (appropriately measured) should probably be included in market simulations, as the None alternative can source share differently from the brands within the simulation scenario.

The Voice of the Patient (Christopher Saigal and Ely Dahan, UCLA): Conjoint analysis may be used to assist doctors in understanding their patients' preferences for different approaches to treatment when facing serious health issues such as prostate cancer. One of the challenges is to develop the attribute list, as the issues and terminology that doctors prominently consider do not necessarily align with the way patients think about the options and the words to describe them. Dr. Saigal showed how attributes and levels were selected for a recent conjoint analysis study regarding patient preferences for treatment of prostate cancer. Qualitative interviews were conducted with 17 patients to learn the issues and language used regarding treatment options, side effects, and outcomes. This led to about 1,000 total parsed quotations. A research team was able to identify 15 themes from these quotations. The statements were printed on cards, which were in turn rated in terms of similarity, and then an agglomerative clustering algorithm was used to identify key attributes to use in the conjoint analysis. Four of the resulting attributes were already quite obvious to the researchers. Three attributes weren't so obvious, but were common themes raised by the 17 patients used in the qualitative phase. An adaptive, CBC-like conjoint survey was used that could report the results to the patient and his doctor in real time. The conjoint survey was compared to other standard survey approaches the healthcare community uses for such research (time trade off, and ratings), and was found to be just as usable and doable from a patient standpoint.

An Overview of the Design of Stated Choice Experiments (Warren F. Kuhfeld, SAS Institute Inc. and John C. Wurst, The University of Georgia, and Atlanta Marketing Sciences Consulting, Inc.): Warren reviewed the principles and goals of experimental design for CBC studies. Experimental design involves selecting combinations of profiles in sets that minimize the error in the beta estimates. Experimental designs can involve generic attributes (that apply across alternatives) or alternative-specific (applying exclusively within the alternative). One of the common measures of efficiency (D-efficiency) has the complication that the preference weights (the betas) enter into the computation of efficiency. Thus, one cannot know the D-efficiency of the CBC design unless one first knows the utilities; which presents a chicken-and-egg problem. That said, researchers can come up with reasonable *a priori* estimates that assist them in developing choice tasks that aren't overly dominated in terms of very large utility differences among alternatives. Warren demonstrated the difference between overlapping and non-overlapping levels within choice sets. Partial-profile designs were also described, as a possible solution to the issue of too many attributes. MaxDiff designs were described as balanced incomplete block designs, where balance can be sacrificed in favor of realism. Different approaches to experimental design were contrasted, including contrasts in available software. Imposing certain restrictions upon a design (prohibited combinations of levels, or desired level of level overlap) lead to slight reductions in traditional efficiency measures, but Warren asserted that the overall questionnaire may be more effective and realistic when dealing with human respondents and addressing client needs.

In Defense of Imperfect Experimental Designs: Statistical Efficiency and Measurement Error in Choice-Format Conjoint Analysis (F. Reed Johnson, Jui-Chen Yang, and Ateesha F. Mohamed, Research Triangle Institute): In his presentation, Reed emphasized that developing an effective experimental design involves more than just selecting a design that leads to the most optimal statistical efficiency. Experimental design for stated preference choice studies necessarily involves human subjects, and the most statistically efficient designs are often at odds with the notions of response efficiency. CBC designs can actually be more effective overall if some degree of statistical efficiency is traded off in favor of gains in task simplicity and task realism. Reed also emphasized that rather than asking “How big does the sample need to be?” researchers should be asking, “What’s the right sample size?” The answer to that depends on issues other than statistics, particularly the research budget. Other important issues are the desired precision, representativeness, and the need for segment analysis. Reed did a meta analysis of 31 CBC projects conducted at RTI, to discover what inputs to experimental design had the largest impact upon the quality of the utility estimates. A bootstrap sampling approach was taken to repeatedly select replicate sub-samples featuring different sample sizes. Differences between the replicates on estimated parameters were observed. Sample size by far had the largest impact on the quality of the utility estimates. Other issues related to optimizing the study design (D-efficiency) had much lower impact than practical issues such as number of attributes, number of levels, number of tasks per respondent, and especially sample size. Reed concluded that research budget is better spent on these practical issues (especially sample size) rather than spending extra money optimizing the experimental design to obtain slightly higher D-efficiency.

CBC Design for Practitioners: What Matters Most (Joel Huber, Duke University): Joel explained that the major drive over the years to increase the statistical efficiency of CBC questionnaires has in turn made them harder for respondents. But, the movement away from small (especially single-block) main-effects designs toward larger designs with many blocks that support higher-order interactions effects has been particularly valuable. Joel emphasized testing designs prior to going to field, using synthetic respondents, to assess how well these designs can estimate the parameters of interest, including interaction terms. He also recommended analyzing the data by different blocks, to test for a block effect. Joel explained that the traditional measure of D-efficiency assumes respondents answer questionnaires much like machines would: that the error in their choices does not depend on the complexity of the task they are given. But this is hardly the case. Designs that are more D-efficient (greater utility balance, more attributes per alternative, more alternatives per task) usually lead to increased respondent error and simplification strategies. Joel suggested that partial-profile designs may be helpful in particularly challenging choice situations, when respondents face especially conflicting tradeoffs. But, he also acknowledged that partial-profile designs may encourage respondents to focus attention on less important attributes. Although the marketing research community has by and large embraced triples and quads within CBC interviewing, Joel suggested that there are contexts (particularly with challenging health care tradeoff studies) where pairs may work better. But, he warned that pairs may lead to simplification heuristics, such as the majority of confirming attributes decision rule. He recommended that researchers add complete overlap for a few attributes in each choice task, starting with many attributes, and graduating to fewer attributes with complete overlap in later tasks. And, importantly, one should construct attributes, levels, and choice tasks to mimic reality as much as possible.

Adaptive Best-worst Conjoint (ABC) Analysis (Ely Dahan, UCLA Medical School): Most of the conjoint research done today involves collecting data from a sample, and estimating utilities once fielding is complete. Ely described a different situation, where the focus of analysis was on each individual (rather than summaries of respondent groups). The particular context was doctor-patient communication regarding options for treating prostate cancer. Respondents would complete a short conjoint survey, after which a report would be produced in real time that quantified the respondent's preferences and allowed the doctor to gain better insight into the particular needs and sensitivities of that specific patient. Ely described a clever adaptive CBC procedure that employed OLS estimation from best-worst concept choices within quads. Four treatment options (with associated outcomes) were presented to a patient, and the patient marked which was the best of the four, and which was the worst. The adaptive approach constructed later quads based on responses to the earlier choice tasks, and the interview terminated once enough information was gathered to permit robust individual-level estimates via OLS. Ely compared the results of the conjoint exercise to those from two other techniques that are often used in the healthcare community to assess patient preferences (time trade off and ratings tasks). He found the conjoint approach to be more effective in predicting holdout judgments. He also compared his OLS approach to CBC/HB estimation across the sample, and found the OLS results to be nearly as effective at predicting holdouts. However, because real time individual results were required for this case, CBC/HB analysis on the sample would not be feasible for his particular needs.

Maximizing Purchase Conversion by Minimizing Choice Deferral: Examining the Impact of Choice Set Design on Preference for the No-Choice Alternative (Jeffrey Dotson, Vanderbilt University, Jeff Larson, Brigham Young University, and Mark Ratchford, Vanderbilt University): It's well-known in the literature that some choice contexts make it more likely that respondents will select the no-choice alternative (walk away from the choice without buying anything). For example, options that are utility balanced make for more challenging decisions, and increase the likelihood of the no-choice alternative. If one option is clearly dominated by another option, the likelihood of no-choice decreases. The data set used was one from an online travel website, where 15,000 searches had resulted in only 166 actual purchases (a conversion rate of less than 1%, or a 99% of no-choice). When the goal of the firm is to reduce the no-choice option (increase buying behavior), a model that examines how the context of choice alternatives affects the outcome is of particular importance to the bottom line. Jeffrey described different mechanisms for incorporating these contextual phenomena into CBC model specification. On one hand, they can be parameterized into the standard logit estimation. But, Jeffrey and his co-authors conceptualized and built a Poisson Race Model of Choice. In many ways, this model is similar to multinomial probit or enhanced randomized first choice. Jeffrey demonstrated that the proposed model has greater model fit. The model also gave insights into how to improve sales on the travel website. Interestingly enough, some of the recommendations of the model are contrary to existing “best” practices for displaying searched options on the website.

Menu-Based Choice Modeling (MBC): An Empirical Comparison of Two Analysis Approaches (Paolo Cordella, Carlo Borghi, Kees van der Wagt and Gerard Looschilder, SKIM Group): A new kind of conjoint questionnaire, menu-based choice (MBC), is becoming more popular over the last decade. With Sawtooth Software releasing a new program for MBC analysis, it naturally led Paulo and his co-authors to investigate how well different models for fitting the data work in practice. MBC software’s “serial cross effects” model estimated via HB was compared to a method developed within SKIM, which incorporated elements based on “sampling of alternatives” (as shown by Ben-Akiva and Lerman in their 1985 book), but extended it to include more information per choice task and to employ HB estimation. The authors found that both methods produced nearly identical holdout prediction results. In terms of aggregate choice probability across 12 menu items and 2 holdout scenarios, the aggregate R-squared of prediction was 0.99. And, both methods were able to predict individual-level holdout combinatorial choices with an accuracy of about 40% (meaning the *exact* combination of 12 options from the menu was predicted correctly for the holdouts). Paulo recommended that practitioners utilize the MBC software, as the data processing and model building is faster than with SKIM’s approach. However, the use of MBC’s “serial cross effects approach” means that users need to take time to consider which cross effects should be included in the model, which involves some complication and a learning curve.

Building Expandable Volume Consumption onto a Share Only MNL Model (Rohit Pandey, Columbia University, John Wagner, and Robyn Knappenberger, Nielsen): Many researchers have been extending the use of CBC into volumetric responses, where respondents indicate not only which products they would select within the task, but a volume of purchase for each. John reviewed some of the approaches for modeling volume within CBC tasks presented at previous Sawtooth Software conferences. Then, he explained an extension of those models that he is using at Nielsen. A major challenge faced with the data set they presented was the extreme variance in the volumes used by respondents, and the number of negative slopes they found with individual-level HB, when fitting concept utility as a predictor of volume. The recommended model they ended up employing started with the “lost volume” approach (where volumes within each task are scaled relative to the maximum volume observed across all tasks for each respondent), and any lost volume within a task is assigned to an artificial None category. However, additional work needed to be done. Rather than accounting for the “lost volume” within the CBC/HB estimation, they dealt with lost volume in a later step. The maximum volume per person was adjusted (smoothed) individually, to lead to the best fit. This approach was free of all potential inconsistencies inherent in the other approaches they tested and the volume sourcing was consistent with the share model.

Creating Targeted Initial Populations for Genetic Product Searches (Scott Ferguson, Callaway Turner, Garrett Foster, North Carolina State University, Joseph Donndelinger, and Mark Beltramo, General Motors Research and Development): Conjoint analysis has been a good tool for helping manufacturers design effective products and product lines. The product line optimization problem, especially for vehicle manufacturers, is quite complex. Not only is the design space huge, due to the number of vehicles in the manufacturer’s line and the number of attributes in the studies, but many of the product alternatives that could be constructed are just not very feasible or desirable. Because exhaustive search is not feasible when facing billions of possible permutations, genetic algorithms (GAs) have been proposed to find near-optimal solutions. Scott compared the standard GA algorithm available within Sawtooth Software's ASM tool to a modified approach that used individual-level utilities to first find "ideal" products (product lines) for each respondent. When the customized (ideal) products are used to seed the first generation of the GA algorithm, the algorithm is able to find a more optimal solution in fewer iterations than the GA that starts with random products.

Can We Improve CBC Questionnaires with Strategically-Placed Level Overlap and Appropriate "Screening Rule" Questions? (Kevin Lattery, Maritz Research, and Bryan Orme, Sawtooth Software): It is well known that many respondents use non-compensatory rules (such as avoiding "unacceptable" levels in conjoint analysis choices). How to ask respondents which levels are unacceptable has been a challenge over the years, as it is known that respondents tend to mark too many levels as completely unacceptable when in reality they are just undesirable. Furthermore, previous research presented at the Sawtooth Software conference has suggested that if respondents tend to screen on certain attributes (in terms of non-compensatory behavior), then increasing the amount of level overlap in the design for those attributes can lead to improvements in utility estimates. Kevin and Bryan did an empirical comparison of different methods to probe "unacceptables" information and add level overlap within CBC designs. Two of the approaches used *a priori* information (from a small pilot study) regarding screening behavior to decide for which attributes to increase the overlap, and two degrees of level overlap were created using fixed design plans. The third method used an adaptive approach to probe if levels that were being avoided in observed choices within CBC tasks were really unacceptable to the respondent. If respondents indicated these levels were unacceptable, then the adaptive algorithm deleted these unacceptable levels from future choice tasks (thus dynamically adding more level overlap on such attributes). A holdout cell of respondents was also included in the study, for out-of-sample validation. There weren't large differences in the performance of these three approaches to experimental design. And, it was apparent the either direct elicitation of unacceptable levels or an adaptive approach still led to over-statement of unacceptable levels. However, the authors found that including individual-level ratings information regarding the levels could improve not only the internal hit rate, but the out-of-sample predictions of holdout shares of preference. Importantly, the authors did not treat the unacceptable levels as truly unacceptable (utility of negative infinity), but treated them as inferior in utility to the acceptable levels.

Being Creative in the Design: Performance of Hierarchical Bayes with Sparse Information Matrix (Jay Weiner, Ipsos Media-CT, and Marcos Sanches, Ipsos Reid): There are many situations that practitioners face where the client needs mean that we can no longer adhere to what is considered best-practices. Marcos presented a case study for a packaged-goods situation in which certain brands only had certain package types. The attributes could not be treated as independent, and too many prohibitions between brand and package size would lead to a deficient design. So, they collapsed the two factors into a single factor with 30 levels representing all non-prohibited combinations of brand and package size. Respondents completed 12 choice tasks, where each choice task included 30 alternatives. Aggregate logit was compared to HB estimation, and either approach led to good predictions of actual market shares (though HB was directionally slightly better). Even when the number of tasks was artificially reduced to 5 tasks, the HB model still led to reasonable predictions of market shares, showing it to be robust even in quite sparse data situations. Because aggregate logit produced results nearly as good as

HB in this condition, Marcos suggested it should be an option when data conditions become especially sparse.

Leveraging the Upper Level Models in HB for Integrated Stakeholder Modeling (Brian Griner, Quintiles Consulting, and Ian McKinnon, Kantar Health, Pieter Sheth-Voss, Proven, Inc., and Jeffrey Niemira, Quintiles Consulting): In some types of markets, such as for pharmaceuticals, there are multiple stakeholders whose decision affect the overall success of the drug in the marketplace. Models that incorporate their relevant attributes and preferences can be helpful to better manage all stakeholder relationships to improve the likelihood of success. Brian reviewed the main steps in Linked Modeling: 1) Identify key stakeholder groups, 2) Outline conceptual model and key linking attributes, 3) Create separate designs for each group including linking variables from the other groups, 4) Estimate separate models per group, and 5) Link models in the simulator. He described the use of the model where multiple stakeholders (doctors, patients, and payers) are involved. Importantly, Brian described how he was able to append all three CBC datasets into a single one, with covariates for estimation under CBC/HB. The results of all models were built into a single Excel-based simulator. Because the share of preference outputs from one stakeholder group are used as inputs to the other models, the re-calculate button has to be cycled multiple times until the predictions for all three stakeholder groups stabilize. Using the model, the attributes that make the most impact on the success of the drug can be tested through sensitivity analysis, leading to better overall strategy.

Modifying Bayesian Networks for Key Drivers Analysis: An Overview of Practical Improvements (Mike Egner and Robert A. Hart, Jr., Ipsos Science Center): The authors argued that Bayesian Networks generally provide a superior method for estimating the effect that certain attributes have upon outcomes such as overall satisfaction or loyalty (key drivers analysis). Bayesian Networks do better than most techniques the authors have examined at handling the common problem of multicolinearity. They have come to that conclusion based on simulations, where they have measured the ability of different methods to recover true (known) drivers that affect the outcome variable. This was largely due to ability of Bayesian Networks to model the structural relationships between drivers rather than assuming they don't exist, or arbitrarily splitting importances across correlated drivers. However, the authors have had to make a few modifications to Bayesian Networks to achieve these positive outcomes. They used a tree-based cell aggregation technique to work around the problems of small cell sizes. Furthermore, they incorporated causal structural search algorithms to identify the direction of causal relationships. They also bootstrapped the data, and averaged the results across the bootstrap iterations. Finally, a simulation approach was used to assess the impact of variables upon the outcome, rather than just looking at average coefficients.

*** Volumetric Conjoint Analysis Incorporating Fixed Costs** (John Howell and Greg Allenby, Ohio State University): There are many products purchased today that involve fixed and variable costs. A classic example is an ink jet printer, where the price of the printer itself only makes up a small portion of the total costs of ownership over time, due to the significant expense involved in buying ink cartridges. John described a case study for an agricultural test machine that involved a single selection of a reader machine (fixed cost), and volumetric indication of number of test strips purchased (variable cost). Since the number of units of test strips purchased in each choice task was given, traditional CBC analysis was not an option. John presented an HB implementation of an economic model of choice, and compared the results to traditional CBC analysis with a simple volumetric extension to account for number of test strips.

He found that the proposed model performed better when fitting holdouts than the traditional CBC with a volumetric extension.

A Comparison of Auto-Recognition Techniques for Topics and Tones (Kurt A. Pflughoeft, Maritz Research and Felix Flory, evolve24): Machine-based approaches for analyzing open-ended text are becoming widespread over the last decade. These approaches can parse text and score comments in terms of topics, as well as sentiment (positive, negative, or neutral comment). Proprietary commercial software solutions are available, such as Clarabridge and Lexalytics. Another software is freely available within R (inverse regression). Kurt reported on an empirical test involving open-end comments about hotel stays. 1000 human-scored sentences were compared to the evaluations of the Clarabridge, Lexalytics, and Inverse Regression approaches. 14,000 sentences were used to train the Inverse Regression model. Clarabridge has already been trained over many years, specifically regarding hotel stay comments. Lexalytics was not specifically trained on hotel comments, but relies on a generic training involving many product categories. The results showed that the Clarabridge approach was most closely aligned with human coding results, followed by Lexalytics, and the Inverse Regression approach. But, all three methods were fairly comparable, and showed a high degree of agreement. Kurt acknowledged that the open-end comments used in this case were quite clean, but it becomes harder for analytical approaches to score as well in the presence of sarcasm, abbreviated words, and misspellings.

(* Recipient of best-presentation award, as voted by conference attendees.)

GAME THEORY AND CONJOINT ANALYSIS: USING CHOICE DATA FOR STRATEGIC DECISIONS

CHRISTOPHER N. CHAPMAN
GOOGLE[1]
EDWIN LOVE
WESTERN WASHINGTON UNIVERSITY

ABSTRACT

We demonstrate an approach to combine choice-based conjoint (CBC) market simulations with strategic analysis using game theory (GT) models. This applied case builds on the limited prior research that has explored the use of game theory and conjoint in combination. We argue that this approach helps to clarify strategic decisions, to communicate CBC results more effectively, and to focus research more clearly on business goals. We present two cases of GT+CBC in a retail product line, and discuss considerations to conduct such analysis properly. We provide an introduction to GT+CBC for strategic marketing applications and describe how it was used successfully in two industry cases.

INTRODUCTION

In practitioner settings, presenting the results of choice-based conjoint analysis (CBC) is complex due to the large number of possible analyses and the mass of data available. Although an analyst may wish to represent the breadth and depth of results, stakeholders and decision makers may be distracted or led astray when they focus on details. For instance, in our experience stakeholders often inspect granular effects, such as "Is feature 3 really preferred over feature 5?" As experienced practitioners know, in random utility models, details such as point estimates may be unreliable and do not present the complete information that is available, such as the distribution of individual parameters. This means that inspection of such estimates can be misleading.

More importantly, such details of a CBC model are often distant from the actual decisions at hand. We propose that executive stakeholders generally should pay little attention to the specifics of CBC analyses, and instead should spend time considering evidence that more directly informs the strategic decisions at hand. Those may be questions such as: "Should we invest in feature X? What will happen to our brand if we undertake action Y?"

Game theory (GT) presents an approach to address such decision making in the face of uncertainty. If one is able to (1) model possible decisions (business actions), (2) model potential

[1] Current affiliation. The research described here was performed while the first author was affiliated with Microsoft and is presented with permission.

competitive and market responses to those actions, and (3) assign outcome metrics, then GT potentially can assess likelihood of those outcomes and the forecasted effect of the decisions (Myerson, 1991). Prior research on GT with conjoint has been limited (cf. Choi and Desarbo 1993; Smith 2000) and not yet widely known. We hope that the cases presented here will advance game theory in marketing research generally, and especially in relation to current developments in CBC.

The quality of a GT analysis is dependent on the accuracy of the strategic model and the data that informs it. However, it is possible to model uncertainty in assumptions and to examine the estimated outcome of a decision under various assumed conditions. Thus, even in cases where outcomes are unclear, one may use GT to examine the likelihood across multiple sets of assumptions.

Consider one possibility: across a range of possible conditions and models, GT indicates that an identical decision is warranted. In such a case, one will feel substantially more confident in recommending the strategy. A contrary possibility is when GT shows that expected outcomes diverge on the basis of minor differences in assumptions and actions. This is also valuable to know because it suggests that a decision is highly uncertain or that we need more information. In either case, GT is a valuable adjunct to strategic analysis.

We outline here an approach to present CBC results when specific decisions are needed in the context of market competition: to combine CBC market simulation with game theory (GT) models. We argue that focusing on a strategic decision can make CBC results more useful in a business context. The GT literature is too extensive to review in depth in this paper. However, we hope that the cases here provide a useful introduction to GT, are illustrative for CBC practitioners, and will inspire broader consideration of GT models among conjoint analysts.

Pairing Game Theory with CBC

There are seven steps required to create a game theory model informed by CBC. First, one must ***identify the business decision(s)*** under consideration and the associated outcome metrics of interest. For example, one might consider the decision of whether to enter a new foreign market, with an outcome metric of expected profit over some period of time.

Second, one must ***identify potential competitive responses*** to the decision point. In the case of new market entry, a competitive response might be that existing participants in the market would cut their prices in an attempt to retain share and make entry unprofitable.

Third, one must ***consider responses "by nature"*** that occur according to outside forces or chance and are not determinable by any participants. An example of this would be refusal of consumers in the new market to purchase goods coming from a foreign entrant. In game theory jargon, such possibilities are likely to involve "information sets" where a branching condition is not knowable in advance but becomes known at some later point.

Fourth, one must ***assess the various outcomes*** of other participants and the data needed to inform all of those. For instance, in the foreign market scenario, although the new entrant might be interested in profit, a current participant may be primarily interested in retaining share. It is not required that players have identical goals; but it is necessary to determine the reasonable competitive decisions that could influence outcomes for the focal player.

Fifth, it is necessary to ***represent these decision points***, inputs, naturalistic paths, and final outcomes as a strategic game, and to ensure that the structure and outcomes are complete. In principle, this is easy to do; however, that belies two points: (a) the game must be plausibly correct so that the relevant factors are modeled; and (b) the stakeholders must be convinced of the structural correctness.

Sixth, one must then ***collect the data*** required to calculate each of the outcome metrics. In the present example, one might collect data from a CBC study to estimate preference share in the new market using a market simulator, which is then matched with the expected cost of goods to determine profit. It is often the case that some information is unknown, especially as regards other players' intentions and the likelihood of natural outcomes. In those cases, one may substitute expert guesses where better estimates are unavailable. It is advisable to run multiple simulations using both optimistic and pessimistic estimates to determine a likely range of results.

Seventh, the game may then be subjected to ***estimation*** using various methods. The most common way to analyze a game is to search for Nash equilibria, which are points where players' strategies are stable, given the strategies of other players. In a Nash equilibrium (NE), there is no decision that any single player could take to obtain a better result, given the strategy adopted by other players, and this is true for all players; thus, these are particularly plausible outcomes if players are informed and rational. A particular game may have no NE, or it might have one, or many.

Importantly, an NE does *not* imply a global optimal outcome for any single player or for all players. It merely implies a likely and stable outcome, even if suboptimal. A well-known example is the “prisoners' dilemma,” in which two assumed criminal partners are separately offered a deal to testify against the other: testify and you'll receive between zero to two years in jail; but if you refuse and the other person testifies against you, you'll receive 10 years. Assume that the payoff for each participant looks like the following (for grammatical convenience, we'll assume both players are men):

You testify, he testifies:	2 year in jail for you	(2 for him)
You don't testify, he testifies:	10 years in jail for you	(0 for him)
You testify, he doesn't:	0 years in jail for you	(10 for him)
You don't testify, he doesn't:	1 year in jail for you	(1 for him)

This may be shown as a “strategic form” matrix as shown in Figure 1, where the numbers in parentheses indicate the outcomes for *(Player 1, Player 2).*

Figure 1: Years in Jail as a Result of Testifying or Not
(Prisoners' Dilemma, Strategic Form Payoff Matrix)

		Player 2	
		Testify	Not testify
Player 1	Testify	(2, 2)	(0, 10)
	Not testify	(10, 0)	(1, 1)

Assume that you are Player 1 and the other person is Player 2. If Player 2 testifies (the first column above), then you are better off testifying (receiving 2 year instead 10 years). If he doesn't testify (the second column above),then you are still better off testifying (0 years vs. 1 year). Thus, regardless of what the other player decides, your best strategy is to testify. Because the payoffs are symmetric, the same analysis applies to Player 2.

In this case, testifying is a stable strategy and NE, because given the other player's decision, there is no way to achieve a better outcome. The expected outcome, then, is for both of you to testify and serve 2 years in jail. Although this is a stable and rational strategy under the assumptions here, it is globally suboptimal to the case in which both players refuse to testify and therefore end up with only one year in jail. The problem with the cooperative ("not testify") strategy is that neither player has an incentive to follow it: if you truly believe that the other person is *not* going to testify, then it is in your interest to testify and walk away with no jail time.

There are various ways to solve for NE. For simple games, as in the prisoner's dilemma and Case 1 below, it may be possible to evaluate them by hand by considering the dominant outcomes. For more complex games, as in Case 2 below, it may be possible to find an analytic solution through computational processes such as solving simultaneous equations (McKelvey, McLennan, and Turocy, 2007). In other complex cases, solutions expressed as long-run expected outcomes might be approximated through repeated simulations (Gintis, 2000).

There is no general assurance of finding a solution or being able to estimate stable outcomes; an attempt to create a game that models a complex situation may result in a non-tractable model. To increase the odds of finding a solution, one should keep a game as simple as possible in a given circumstance.

The primary results of finding equilibria are estimates of the probabilities that the game concludes at each outcome state. In the prisoners' game above, there is one NE with an outcome node with probability 1.0: both players testify. In more complex cases, a successful solution will yield the probability for each outcome state. This may be interpreted as the frequency that one would expect to end up in that state, if the game were played an infinite number of times with the same assumptions.

An overall expected value for the game may be computed by vector multiplication of probabilities by outcome value; a credible interval may be constructed by ranking the values weighted by probability. For instance, suppose a game has three outcomes with *(value, probability)* = (1, 0.02), (100, 0.96), (10000, 0.02). The expected value across infinite iterations is 1*0.02 + 100*0.96 + 10000*0.02 = 200.96, while the 95% credible interval is [100, 100].

A few other notes are in order. Estimation approaches other than evaluating NE are available and can be useful in situations in which equilibria do not exist or are doubted to be applicable. In general, we recommend to evaluate for NE first, and to consider other methods after NE feasibility is determined. Repeated games and cooperative games may lead to solutions that are quite different from NE. In the game above, if the players above are 100% cooperative, then the dominant strategy would be *not* to testify.

In this paper, we consider non-cooperative games because those are often typical of competitors in a market. Additionally, our cases concern two-player, non-repeated games; for more complex situations than those described here, one will wish to review the GT literature to determine which models may apply. An extensive GT literature exists in various fields, including economics (e.g., Dutta, 1999), psychology (e.g., Gintis, 2009), and evolutionary biology (e.g., Maynard Smith, 1982).

Case 1: Whether to Develop a New Feature

The manufacturer of a PC accessory hardware device was considering the question of whether to add a feature (feature X) to its product line, after learning that Feature X was going to be available from component suppliers in the near future.

For our purposes here, the category and feature are disguised. However, the feature X component was somewhat analogous to a higher-speed processor in a computer: it would bring a higher “spec” to the products and was seen as desirable by consumers. On the other hand, it would add cost and might not make much, if any, difference in users' actual experience. Importantly from a GT modeling perspective, this product category had two dominant players (the manufacturer in question and one other firm), and feature X would be available to both.

Various business stakeholders had differing opinions about whether feature X should be included. Some thought it would appeal to customers and grow the overall size of the category (relative to other products), while others argued that it would simply add cost and make the category less profitable. The latter group was, in effect, arguing that this was a variety of the prisoner's dilemma described above: adding the feature (e.g., “testifying”) would benefit either player in isolation because it appealed to consumers and would take market share from the other player; yet if both players added it, they would simply be dividing the same market at a higher cost, and thus both would be worse off.

The authors realized that this was an almost canonical case of a two-player strategic game: two players faced a single and identical business question, and they had to act simultaneously without knowledge of the other's intent. Additionally, the authors could model the various possible outcomes of the game: because they had fielded more than a dozen conjoint analysis studies in this category, they had the data needed to model a likely market outcome. As shown in Chapman et al (2009), conjoint analysis had proven to be a robust and useful indicator of market outcomes in this product category.

Game 1

We modeled Case 1 as a two-player, simultaneous, one-step game with identical goals for the two players. Since each player had two options, to include feature X or not, the game had four possible outcomes – four “strategies” in GT jargon – for (us, them) respectively: (not include, not include), (not include, include), (include, not include), and (include, include). This game is shown schematically in Figure 2.

Product executives identified the division's strategic goal as maximization of market share, specifically to gain share from the other dominant player. This led us to compute the four sets of outcomes for each player as preference share for likely product lines, with and without feature X, according to the scenario. We then computed share for the players in each of the four outcome scenarios using a comprehensive set of CBC data, using Sawtooth Software Randomized First Choice market simulation.

Figure 2: The Game for Case 1
(Preference Share in Strategic Form Payoff Matrix)

		Player 2 (them)	
		Do not include	Include feature
Player 1 (us)	Do not include	(X11, Y11)	(X12, Y21)
	Include feature	(X21, Y12)	(X22, Y22)

NOTHING IS CRUCIAL: WHY OUTSIDE GOOD MUST BE MODELED

So far the situation seems simple: for the game we just need the decision points, and the outcomes measured by CBC market simulation. Right? Not quite; there's a fundamental problem that must be resolved: to estimate what is likely to occur when a product lineup changes, one must have some reference to evaluate the lineup as a whole against other possible alternatives.

Here's the problem: if both players add feature X, and product lines are the same otherwise, then the relative preference on an individual by individual basis should not change. In that case, the outcome of the strategy *(include, include)* would be identical to the outcome of *(don't include, don't include)*. If the latter has share of (40, 60) then the former should also have share (40, 60). Under a strict first choice market simulation model, that is exactly what we would see. Suppose that for an individual, the utilities for player 1 and player 2, without feature X, are *U1* and *U2*; and the utility of feature X is *X*. Since conjoint assumes that utilities are additive, if *U1=U2* then *(U1+X) = (U2+X)*. Both players have made the same move at the same time, so their relative positions are unchanged.[2] In the strategic game shown in Figure 2, this would imply X11 = X22 and Y11 = Y22. In other words, in this model it makes no difference what the two players do, as long as they both do the same thing. This is obviously not a useful or correct model.

The key to addressing this problem is to have a reasonable baseline from which one can determine whole-market change and not simply relative preference. This is the problem of reference to outside good, in economic jargon, or the "None" choice in conjoint terms, and is a complex area with ongoing research (cf. Dotson et al, 2012; Howell and Allenby, 2012; Karty, 2012, all in this volume).

Here, we note several ways to address the None issue in routine conjoint models:

1. use first choice preference models in market simulation that are resilient to some issues with logit-model preference estimation (cf. Footnote 2)
2. include a dual-response none option in CBC surveys

[2] For completeness, we should mention the *Red Bus/Blue Bus* problem where multinomial logit share models *do* show a change when the same utility is added to both sides of a relation – but that change is inappropriate and an artifact of the models, not a reasonable market expectation. A full discussion of that phenomenon is available in papers from previous Sawtooth Software Conferences and the AMA ART Forum (e.g., Dotson et al, 2010).

3. when estimating part-worths and running simulations, model expected interaction effects (such as brand*price interaction, which is common in many categories)
4. in product line simulations, include some fixed products that are the same across all scenarios

In Case 1 we adopted all four of those strategies to account for the outside good in simulation and ensure that results were not predetermined to be identical. As a general rule, one should have prior confidence in the simulation models on the basis of research experience in the category.

Another option we did not employ but may be considered is to use hierarchical Bayes (HB) draws in simulations instead of respondent mean betas. Any single draw would have the same issue of preference that is intransitive to feature addition as does any other single-point preference estimate. However, by using multiple draws, such as 1000 estimates per respondent, one obtains a picture of the per-respondent distribution of utility overall. By having a more accurate overall estimate, one could expect that concerns about other sources of estimation bias would be partially mitigated.

A final issue concerns source of volume modeling: feature X may have different effects on the share of products according to how similar or dissimilar those products are. For instance, adding a 50 MPG engine to an automotive line should have a different effect if the line currently has no car above 30 MPG than it would if the line already includes cars with 45 MPG and 60 MPG. Similarity between products ought to inform market simulation, but that is not the case in all simulation methods, such as simple logit model estimation. A strategy that at least partially mitigates the problem is a randomized first-choice model (which we employed); more complex models are available when this is a crucial issue for the category in question (cf. Dotson et al, 2010).

Case 1 Result

Existing CBC data was used to estimate preference shares as noted above. The CBC exercise comprised 8 attributes with 30 total levels, 12 tasks with no holdouts, collected from N=359 respondents, and utilities estimated with Sawtooth Software CBC/HB. Figure 3 shows the result of four market simulations required for the game and performed with Sawtooth Software SMRT randomized first choice simulation. Each outcome listed in Figure 3 shows the estimated sum preference of the modeled product lines for (Player 1, Player 2), including the “none” option in simulation.

Figure 3: Estimated Market Results for Case 1
(Estimated share percentages for *(Player 1, Player 2)*)

		Player 2 (them)	
		Do not include	Include feature
Player 1 (us)	Do not include	(23, 44)	(10, 72)
	Include feature	(61, 20)	(29, 54)

There is one NE that shows strict dominance: *both players include Feature X*.

We can find the NE as follows. First, look at the outcomes for Player 2 in Figure 3. In both rows, the outcome for Player 2 is better if they include the feature (72 > 44, and 54 > 20). Thus, Player 2 should include the feature no matter what Player 1 does. Thus, Player 1 only needs to consider the alternatives shown in column 2, where Player 2 includes feature X. Looking at that column, Player 1 is better choosing to include feature X (29 > 10). Thus, the strategy of *(Include, Include)* is a stable strategy and reflects the rational choice for each player given the game assumptions. The same conclusion would be reached by starting with Player 1 instead.

We note also in Figure 3 that both players see an increase in preference share by following the *(Include, Include)* strategy. In other words, by including feature X, it is estimated that the overall market will grow, with both players taking share from the "None" option (i.e., the share of preference that goes to other brands or outside choices). This indicates that the game as modeled is not a prisoners' dilemma.

We shared these results with executive management, along with the following analysis:

1. we should expect Player 2 to include feature X
2. if in fact they do not include feature X, it is an even stronger reason to include it because of the share we would gain
3. feature X might grow the overall category, which would mitigate concern about increased per-unit cost

Management was convinced by this analysis and ultimately included feature X in the product line. It turned out that Player 2 had not anticipated this, but belatedly added X to their product line, and that sluggishness was somewhat detrimental to their brand image.

At the time of this writing, Player 1 and Player 2 produce all products in the top 10 by unit sales in this category. Eight of those 10 products include feature X.

In short, the analysis here appears to have been correct for determining the market demand for feature X; it helped management to make that choice; and it assisted the product team to advance the brand in the face of a sluggish competitor.

Without the game model that convinced business stakeholders, it is likely that neither Player 1 nor Player 2 would have introduced feature X for one or more years, and would have lost an opportunity to advance their product lines and to meet consumer demand. In a worst-case scenario, this would have risked incursion in the category by another brand if one had discovered the unmet need, damaging the positions of both players. Instead, the category was improved by players 1 and 2 strengthening their positions and delivering better, more highly desired products.

Case 2

Following the success in Case 1, we tackled a more complex problem: could we estimate the likely outcome if an entire product line attempted to redefine its brand position? This would include actions such as redesigning products in terms of features and industrial design, changing the line size and composition, producing new packaging and brand materials, advocating different retail presentation, and adopting a new messaging strategy, all in order to change the line's perception and positioning *vis-a-vis* competitors in the category.

There are many variables in such a question, so we worked with stakeholders to identify a narrower set of questions:

1. if we repositioned as planned, what would be the best-guess outcome for full-line preference share?
2. how much potential downside risk would there be?
3. how would our primary competitor be most likely to respond?

(Due to the sensitive nature of the decisions in this case, we disguise the product and refer to the category and issues using only general terms.)

Important background information for this case is that the category had two dominant players (players 1 and 2) along with several smaller players. There was no question of *how* to reposition; the possible direction had been established by other work, and the question was whether it was likely to succeed. Further, it was known that players 1 and 2 had very similar brand perceptions at the present time.

Unfortunately, circumstances required us to produce an answer quickly, without time to collect new data: executives asked for a best-guess assessment in a few days' time. Luckily, we had data on hand from a recent study in which the category was assessed using a CBC with attributes that reflected the full product line scope, and we had scale-based adjectival ratings of key brands in the space (e.g., rating of each brand on attributes such as value for the money, trendy design, high performance, and so forth).

That data had already been processed to yield three key sets of measures: (a) CBC/HB utilities for brand and other attributes; (b) segmentation of respondents into two segments that differed substantially in both adjectival ratings and CBC utilities; (c) perceptual dimensions (aka composite product mapping, CPM; Sawtooth Software, 2004) for the brands as reflected by adjectival ratings, for each of the two segments.

To make use of that data, we proceeded under the following assumption: *brand value* as measured in CBC brand utility should be related to a *brand's position in perceptual space* as measured by CPM. In short, with several brands, it would be plausible to regress their CBC brand utilities on CPM product dimensions. If a brand repositioned in the perceptual space, that regression might be used to determine a best-guess new value for CBC brand utility. We assumed that the two segments would have different models according to their differing perceptual and utility measures.

To express that model schematically, we postulated the following (where "~" means "varies as a function of", and "| ..." indicates "for a given ..."):

1. *Brand dimensions* ~ *adjectival ratings* | *segment* (composite product mapping, Sawtooth Software, 2004)
2. *Brand utility* ~ *stated choice observations* (standard CBC model)
3. *Preference share* ~ *brand utility* (and other CBC utilities, in a standard market simulation model)
4. *Updated brand utilities* ~ *new position on brand dimensions* | *segment* (linear model regressing utility on principal brand dimensions)

5. *Updated preference share ~ updated brand utilities* (market simulation with new brand utilities estimated from the regression model)
6. *Full preference model* would express the model above as performed separately for two attitudinal segments, weighted according to estimated segment prevalence

There are several questionable assumptions in that model; perhaps the most dubious is modeling brand utility as a function of perception. Among the objections to such a model are that different brands might each be highly valued exactly *because* of their perceptual differences and thus not fit a regression model; that measurement error in both metrics could make it difficult to find any relationship; and that a relationship would in any case be expected to change precisely because of brand repositioning, invalidating conclusions drawn from a model of previous positioning. However, as we show below, a GT model may partially mitigate such concerns if one incorporates estimates of uncertainty.

The Game Model for Case 2

Case 2 poses a simple strategic question: should Player 1 adopt the repositioning strategy or not? Likewise, we were interested whether Player 2 would be likely to adopt the same strategy. We also considered the case that Player 1 might initially stay, but then decide to reposition later, depending on what player 2 did. Thus, the game was:

Time 1:	Player 1 repositions or not, in a specified direction
Time 1:	Player 2 repositions or not, in the same direction
Time 2:	If Player 1 did not reposition at time 1, but Player 2 did, then Player 1 might reposition or not at time 2

Despite that simplicity, there is another complication: an attempt to reposition might succeed fully, succeed to a partial extent, or fail altogether. Thus, *attempting* to reposition does not lead to a single outcome estimate, but rather opens the possibility of multiple different outcomes that occur, in game theory parlance, “by nature,” i.e., which beyond the control of the player.

Such divergences in possible outcomes should be modeled in the game. In the present case, we modeled four possible outcomes:

1. *Full success*: reposition achieves 100% of the regression model change in brand utility
2. *Partial success*: reposition achieves 25% of the regression model outcome
3. *No change*: reposition results in no change in brand value
4. *Backfire*: reposition results in loss of brand value, -25% of the regression model

Note that the levels of possible outcome (100%, 25%, etc.) are not determinable from data but must be specified by expert opinion. This is a strength of GT models, that they are able to use expert opinion to model multiple scenarios when better information is not available and can simulate the outcomes across multiple sets of assumptions.

The full game model for Case 2 had 41 outcome nodes, as follow:

Player 1 choice	**Player 2 choice**	**Outcome nodes**
Do nothing	Do nothing	1
Do nothing	Reposition (4 outcomes)	4
Reposition (4 outcomes)	Do nothing	4
Reposition (4 outcomes)	Reposition (4 outcomes)	16
Do nothing at Time 1 + Reposition Time 2 (4 outcomes)	Reposition (4 outcomes)	16
Total scenarios		41

Some of those scenarios are effective duplicates (e.g., "*do nothing, do nothing*" has the same effective outcome as "*attempt reposition with no change, attempt reposition with no change*"). However, the odds of the branches occurring could be different and they might occur at different times, when different information is available to the players. We shall skip a detailed exegesis of those combinations, except to say that they should be reflected in the game model with careful attention to timing and the appropriate "information sets" available to the players.

We then specified a range of likelihoods for each of those outcomes, again, based on expert assumption. For instance, one set of likelihoods we assigned was the following:

Full success:	10% chance
Partial success:	50% chance
No change:	30% chance
Backfire:	10% chance

We constructed multiple sets of such outcomes likelihoods, using a conservative set of assumptions that regarded full success as unlikely.

To recap, the model comprised the following elements:

- Predicted change in brand utility based on movement on a perceptual map
- Estimated change in preference share based on those brand utilities (market simulation)
- A game reflecting player strategic decisions and subsequent natural outcomes
- Multiple sets of assumptions about the likelihood of those natural outcomes

With those elements in place, we were then able to run the game multiple times, under the various sets of assumptions about natural outcome likelihood. Unlike Case 1 above, the GT model for Case 2 requires estimation software; we used the open source Gambit program (McKelvey, McLennan, and Turocy, 2007).

Case 2 Result

The model results in two components: (1) preference share for each outcome node, as determined by the market simulation using adjusted brand utilities, and (2) the result of the game model, which specifies the likelihood for that outcome node. Given the values and their likelihood estimates, simple descriptive statistics can be used to determine a median outcome, likelihood intervals, and other distribution characteristics.

Across runs with varying sets of parameters as noted above, the Case 2 model yielded the following overall result:

1. If we repositioned, and given the model assumptions, the expected outcome is a gain in preference share of 6 to 12 points vs. the primary competitor in the category.
2. Downside risk from attempting repositioning appeared to be no more than -1 point change in preference share.
3. The competitor was unlikely to respond with a similar repositioning strategy.

On the basis of this analysis, stakeholders felt confident to adopt the repositioning strategy and we found no reason to object to it. The outcome appeared unlikely to be worse that the current position and offered substantial upside opportunity for positioning with respect to the competitor.

The repositioning strategy was adopted soon after this analysis, which was approximately two years before the time of this writing. Market results are unavailable due to confidentiality, but there is no reason to suspect that the choice was mistaken, and the product line continues to follow the repositioned strategy.

Potential Pitfalls of Game Theory with CBC Data

The validity of a game model depends, of course, on the quality of the outcome estimates for the various strategic branches. There are several well-known issues in CBC modeling that are especially problematic for GT models. Indeed, an advantage of GT modeling is that it forces an analyst to confront those issues with CBC deliberately, whereas in other circumstances it would be possible to ignore them and perhaps reach erroneous conclusions.

Problem 1: Market simulation. To have confidence in the GT mode, the underlying market simulations must be reliable. It is best to do GT+CBC in an area where there is enough history to know that the proper attributes/levels are being tested with conjoint, and that market simulations give reasonably accurate results. An analyst will need to decide case by case when it is better to do GT modeling in the face of imperfect data than to do nothing. In any case, it is strongly recommended to use more advanced market simulation models such as RFC or simulating over HB draws instead of simple logit share models.

Problem 2: None. As we described above, an especially important part of a GT model with market simulations concerns the question of outside good. The “None part worth” may be used to get an estimate of that (as described in Case 1 above), but it must be noted that there are many conceptual, methodological, and empirical problems with None. When None is used, an outcome may be more attractive because of its potential to attract new interest. However, when this involves a new feature, it may also simply reflect error in the None distribution that makes anything new appear erroneously to be attractive (cf. Karty, 2011). Recommendation: if your game requires an estimate of outside good – as most interesting marketing cases do – review the issues and options carefully.

Problem 3: Main effects vs. interaction. For many GT+CBC models, it is important to add an interaction effect between key decision components (such as brand and price). Consider the following situation: Brand A and Brand B could each add feature X, which costs the same for both brands. Assume that utility outcome is estimated from CBC using main effects. If there is a brand*price interaction – for instance, because one brand customarily has had higher price points – then adding the same price utility for feature X in both brands would likely be inappropriate. This does not imply one should estimate all possible interaction terms in CBC, but rather the ones most important for market simulation stability and the GT decision points. Quite often that may be one specific attribute (price, design, or a feature) that interacts with brand. If the decision involves one or two specific attributes, those should be carefully considered for inclusion in interaction effects (interacting with brand, price, or other attributes as relevant).

Problem 4: Similarity and source of volume. There are well-known problems with portfolio comparisons when products are similar (Huber, et al, 1999; Dotson, et al, 2010). If the market simulation model has several products that are closely related or functionally interchangeable, the analyst should consider correction for product similarity. Without correction for similarity, and especially if simple logit models are used, then the market simulation estimates may be substantially in error. At a minimum, we recommend to use randomized first choice (RFC) models; there are other procedures available for cases where similarity is high (Dotson, et al, 2010).

Other Notes about Game Theory

There are a few other important GT considerations that arise in marketing applications. We note them here primarily for awareness of GT options.

Payoff imbalance. Businesses typically view active mistakes as much worse than a passive loss or opportunity cost, and utilities may need to be adjusted to account for risk avoidance and other such behavior (cf. Gourville, 2004). For instance, the outcome metric in a game (such as *share* in the games presented here) might be weighted such that *negative* change is calculated to reflect imbalance as viewed by stakeholders, such as 3x, $-(x^2)$ or some other multiple of its actual value, while positive change is estimated with an unchanged metric.

Iterated or sequenced games. In actual situations, strategic decisions may be sequential or iterated, not following a “simultaneous decision” model. Thus, GT models may need to be iterated, and likely will need to model multiple information sets rather than perfect knowledge (cf. Myerson, 1991). A particularly interesting family of games involves those that are iterated indefinitely. These are often used in evolutionary models, and provide alternative methods to find equilibria and stable points in some classes of games that defy analytic solution (Gintis, 2000).

Multi-player games. Game models can be extended in a straightforward way to situations where there are multiple players (multiple brands, etc.). However, the stability of market simulation results is likely to decline, and very close attention should be paid to brand effects, the impact of distribution, channel, and availability, and so forth. Those effects may not be necessary to model formally as part of the market simulation, but should be considered and included in the interpretation.

Conclusion

Game theory has great potential for analyzing strategic marketing situations. We believe that the examples presented here are illustrative of its power to inform business decisions. In particular, game theory is a valuable adjunct to conjoint analysis for three primary reasons: (1) it directly models strategic decisions, which is intrinsically of interest for businesses; (2) it provides a concise way to frame questions, assumptions, and results, allowing for direct focus on the decisions at hand; (3) it can incorporate uncertainty and assumptions, making it a valuable adjunct to sensitivity analysis and situations where data such as conjoint analysis informs only one part of a complex picture.

Game theory has been little used to date in applied marketing science, but with the growth of available data and computing power to simulate complex models, there are many applications for game theory to inform marketing decisions.

References

1. C.N. Chapman, J.L. Alford, C. Johnson, M. Lahav, and R. Weidemann (2009). Comparing results of CBC and ACBC with real product selection. *Proceedings of the 2009 Sawtooth Software Conference*, Del Ray Beach, FL, March 2009.
2. S.C. Choi and W.S. Desarbo (1993). Game theoretic derivations of competitive strategies in conjoint analysis. *Marketing Letters*, 4 (4), 337-48.
3. J. Dotson, J. Larson, and M. Ratchford (2012). Maximizing purchase conversion by minimizing choice deferral: examining the impact of choice set design on preference for the no-choice alternative. *Proceedings of the 2012 Sawtooth Software Conference*, Orlando, FL [this volume].
4. J. Dotson, P. Lenk, J. Brazell, T. Otter, S. McEachern, and G. Allenby (2010). A probit model with structured covariance for similarity effects and source of volume calculations (Working paper, April 2010, http://ssrn.com/abstract=1396232). Presented at the 2009 Advanced Research Techniques Forum (ART Forum), Whistler, BC, June 2009.
5. P.K. Dutta (1999). *Strategies and Games: Theory and Practice*. MIT Press, Cambridge, MA.
6. H. Gintis (2000). *Game Theory Evolving: A Problem-centered Introduction to Modeling Strategic Behavior*. Princeton Univ. Press, Princeton, NJ.
7. H. Gintis (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton Univ. Press, Princeton, NJ.
8. J. Gourville (2004). Why customers don’t buy: the psychology of new product adoption. Case study series, paper 9-504-056. Harvard Business School, Boston, MA.
9. J. Howell and G. Allenby (2012). Local monopolies arising from fixed costs. *Proceedings of the 2012 Sawtooth Software Conference*, Orlando, FL [this volume].
10. J. Huber, B. Orme, R. Miller (1999). Dealing with product similarity in conjoint simulations. Research paper series. Sawtooth Software, Sequim, WA.

11. K. Karty (2012). Taking nothing seriously: a review of approaches to modeling the "none" option. *Proceedings of the 2012 Sawtooth Software Conference*, Orlando, FL [this volume].
12. K. Karty (2011). Bias in main effects models for line optimization. Presented at the 2011 Advanced Research Techniques Forum (ART Forum), Desert Springs, CA, June 2011.
13. J. Maynard Smith (1982). *Evolution and the Theory of Games*. Cambridge Univ. Press, Cambridge, UK.
14. R.D. McKelvey, A. M. McLennan, and T. L. Turocy (2007). Gambit: software tools for game theory [computer software]. Version 0.2007.12.04. http://www.gambit-project.org.
15. R. Myerson (1991), *Game Theory: Analysis of Conflict*. Harvard Univ. Press, Cambridge, MA.
16. Sawtooth Software (2004). The CPM System for Composite Product Mapping. Technical paper series. Sequim, WA.
17. D.B. Smith, and S.M. Whitlark (2000). Sales forecasting with conjoint analysis. In *Conjoint Measurement: Methods and Applications*, A. Gustafsson, A. Herrmann, and F. Huber (eds). New York: Springer.

CONTRAST EFFECTS ON WILLINGNESS TO PAY: HOW CONJOINT DESIGN AFFECTS ADULT DAY CARE PREFERENCES

DAVID BAKKEN,
MICHAELA GASCON,
AND DAN WASSERMAN
KJT GROUP

Setting prices may be just about the most important thing a company does. Economist Paul Ormerod (author of *Why Most Things Fail*) puts it this way: "If a firm makes a big enough mistake on [pricing], or even persists with perhaps relatively minor mistakes for a sufficiently long period, it will fail."

For most products and services, discerning the customer's "willingness to pay" a particular price varies from merely difficult to almost impossible. For one thing, consumers have a vested interest in gaming the system, so that self-reported willingness to pay is often suspect. For another, consumers often do not know what they are actually willing to pay until they are confronted with a purchase decision. Over the past few decades market researchers have devised a variety of methods, with varying success, for uncovering these demand curves.

MARKET RESEARCH PRICING METHODS

Methods for estimating willingness to pay (WTP) can be grouped according to whether they attempt to elicit the consumer's *reservation price* (the highest price that a customer is willing to pay for a specific value proposition) directly or indirectly and whether they estimate actual or hypothetical WTP. For the most part, direct elicitation methods ask consumers "How much are you willing to pay?" and indirect methods observe consumers' behavior in actual or simulated purchase situations where price can be varied systematically. The "gold standard" for pricing is the *monadic* price experiment, an indirect method which can be conducted in-market (actual WTP) or in a contrived setting (hypothetical WTP). In full expression, the monadic experiment consists of a sample of consumers who are randomly assigned to different *price cells* (experimental treatment conditions) with one (and only one—hence "monadic") of several possible prices presented in each of these cells. For an in-market experiment, these consumers are presented with actual offers at prices determined by the experimental cell they are assigned to. For contrived settings (such as a survey), a sample of consumers may be placed in a simulated purchase situation (with competitive alternatives), or they may simply be asked how likely they are to purchase at the price they see.

Monadic price experiments have a few important drawbacks. In-market experiments are cumbersome, expensive, and may need to run for an extended period of time. Simulated experiments typically require large samples of consumers, making them expensive compared to other methods. For most products and services, *choice-based conjoint* pricing models are seen as the next best (and often better) alternative to monadic experiments. In this context, we can think of choice-based conjoint as a *repeated measures* pricing experiment in a contrived setting. A sample of consumers is presented with a series of simulated purchase situations where price as well as other value-adding attributes or features are systematically varied. One significant advantage of choice-based conjoint over a monadic price experiment is that the conjoint method reveals a response function that encompasses variation in the product or service value proposition as well as variation in price. Conducting a monadic experiment of the same complexity as the average conjoint pricing model would require perhaps hundreds of individual cells and possibly thousands of respondents. (For an evaluation of alternative approaches to measuring WTP, see Miller, Hofstetter, Krohmer, and Zhang, 2011)

Although choice-based conjoint is more efficient, in general, than either in-market or contrived setting price experiments, this approach is more expensive than other market research approaches. Additionally, special statistical analysis tools and a certain level of technical competency are pre-requisites for this method. For these reasons, market researchers sometimes resort to direct elicitation (sometimes referred to as "quick and dirty") pricing methods. Chief among these is the van Westendorp Price Sensitivity Meter (PSM), introduced in 1976 by Dutch economist Peter van Westendorp. The PSM consists of four questions that can be asked in a market research survey. Following exposure to some information about the product or service of interest, consumers are asked:

> At what price would this be *too expensive* and you would not buy it?
>
> At what price would this be *getting expensive* but you would still consider buying it?
>
> At what price would this be a *good value* and you would likely buy it?
>
> At what price would this be *too cheap* and you would doubt its quality?

Question wording varies a bit from one practitioner to the next, and common variations include adding a price range (for example, "At what price between $50 and $250 would this be too expensive?") and adding one or more *purchase intent* questions ("At the price you said would be a 'good value,' how likely are you to buy?").

Unlike choice-based conjoint, there has been little systematic research validating the demand curves produced by this method. Rather, use of the PSM seems to be sustained by high face validity and perhaps by the fact that, for well-established product and service categories, it often returns the current pricing for those categories. In addition, the demand curves estimated by the PSM may be seen as "good enough" for the pricing decisions they are driving.

Rational Pricing Methods, Irrational Consumers

All of these market research pricing methods assume a certain degree of rational decision-making on the part of consumers. Consumers look at the offer, make some assessment of the utility they expect to derive from the purchase, and assign a dollar value to that utility, which is observed either directly or indirectly. In recent years, however, we've seen both empirical and theoretical challenges to this assumption. Notably, Dan Ariely (2008), a behavioral economist, has shown that responses to price can be affected by comparisons to alternatives or even by purely coincidental information. The first effect, which Ariely terms "relativity" but that is more broadly known as the "attraction effect" in decision field theory occurs because we find it easier to compare alternatives that are similar to one another and difficult to compare things that are dissimilar. For example, if you were faced with a choice of dining at two equally expensive restaurants serving very different cuisines which are equally appealing you are likely to have a hard time choosing between the two restaurants. If, however, we add a third alternative—a less expensive restaurant serving the same cuisine as one of the more expensive restaurants—the choice becomes much easier. Ariely's research demonstrates that once this third alternative (in effect, a "decoy") is introduced, most individuals *ignore* the more expensive option serving the different cuisine and make their choice between the two restaurants serving the same cuisine.

Another effect is "arbitrary coherence." Ariely shows that something as simple as making people aware of an unrelated number—in one case, it's the last two digits of a person's social security number—can have a profound impact on what they say they are willing to pay for something like a vintage red wine. The unrelated number appears to function as an unconscious "anchor" when the individual is deciding what to pay. In most pricing situations, the anchor is not some arbitrary number but the price of some other, perhaps only distantly related, product or service. For example, if a well-equipped automobile is priced at $25,000, a $10,000 flat panel television might seem to be overpriced.

Because critical pricing decisions are made on the basis of these market research methods, we wanted to determine if these effects might introduce biases into estimates of willingness to pay obtained from both direct and indirect pricing research methods. We conducted two separate experiments around willingness to pay for adult day care services. In the first experiment we varied the comparison information we provided in a van Westendorp price sensitivity exercise. In the second experiment we extended this to a choice-based conjoint exercise.

Experiment 1:
Effect of price range and reference price comparison on van Westendorp estimates of willingness to pay for adult day care services

Adult day care (ADC) services provide out-of-home health, social, and therapeutic activities for adults with cognitive or functional impairment. ADC is an alternative for some individuals to either assisted-living (residential) facilities or in-home care. ADC is often used to provide periodic relief to informal (i.e., unpaid) caregivers, often a spouse or other family members. A survey of long-term care costs conducted in 2009 by the MetLife Mature Market Institute estimated the nationwide (United States) average daily rate to be $67, with the lowest and highest rates in individual states of $27 (Montgomery, Alabama area) and $150 (Vermont), respectively. By way of comparison, the average for one month of assisted living was $3,130, and the average hourly rate for a home health aide was $21.

We designed a 3 X 3 factorial experiment, varying "price range" information for ADC in the van Westendorp questions and sometimes introducing an "outside good" price reference in the form of either the monthly rate for assisted living or the hourly rate for a home health aide, to test these two hypotheses:

> *H1:* The location of the price range will impact the average willingness to pay as measured using van Westendorp PSM.
>
> *H2:* In the absence of a price range, outside comparisons will impact the average willingness to pay as measured using van Westendorp PSM.

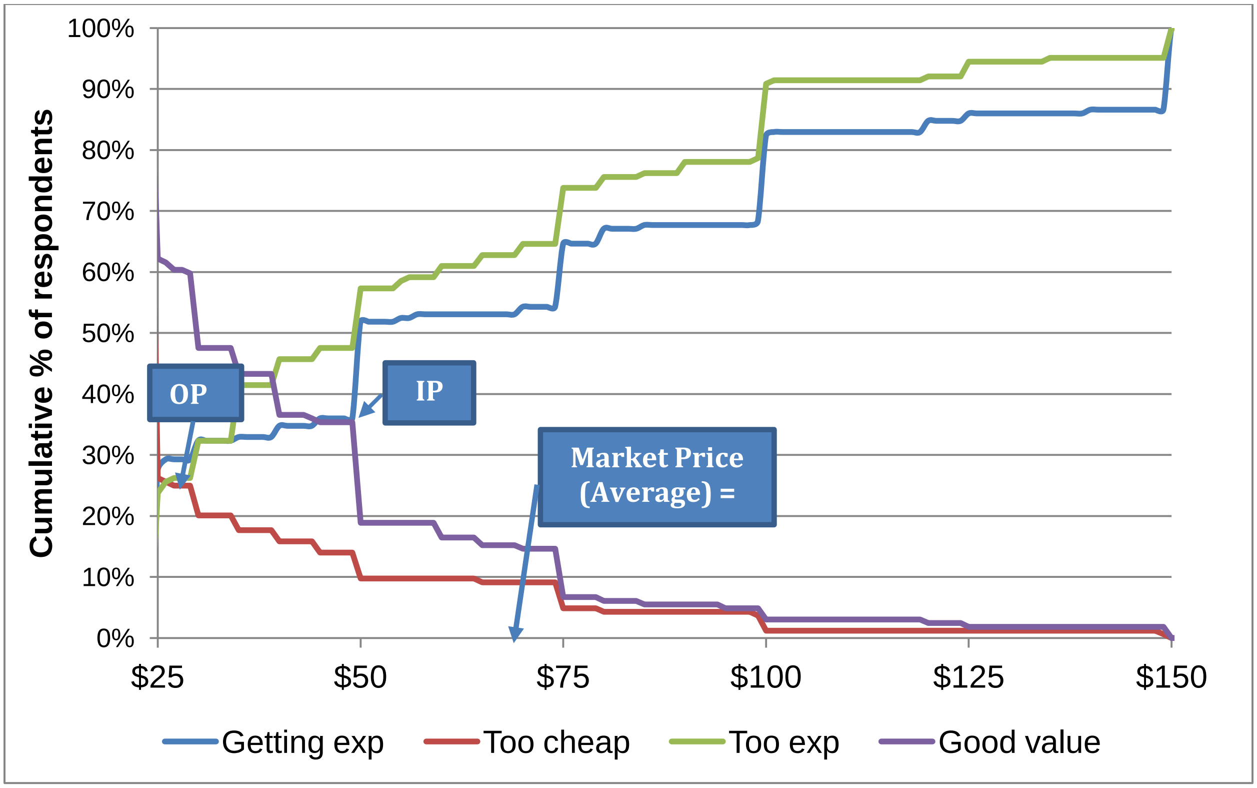
The three levels for the price range factor were: no price range, $25-$150 per day, and $50-$200 per day. The three levels for the outside reference price were: no outside reference price, home health aide (HHA) at $22 per hour, and assisted living (AL) at $3,100 per month.

The experiment was embedded in an online quarterly survey (fielded in December, 2010) that tracks informal caregiving. The 1,487 U.S. adults who completed the survey were assigned randomly to one of nine experimental cells, resulting in about 165 respondents per cell.

Experiment 1 Results

In its classic form, the results from the van Westendorp PSM are analyzed by calculating the cumulative frequencies for each question and finding the four points at which these distributions intersect, as in Figure 1, which shows the results for a single cell in our experiment. The *acceptable* price range is bounded by the intersection of the "getting expensive" and "too cheap" curves (the point of marginal cheapness, or PMC, which is actually below the low end of the price range on this chart) and, on the high end, by the intersection of the "too expensive" and "good value" curves (the point of marginal expensiveness, or PME). Within the acceptable range the point at which the "too expensive" and "too cheap" curves intersect is the optimal price point (OPP), and the point at which the "getting expensive" and "good value" curves intersect is the indifference point (IP). Incidentally, this method applied to these data would lead us to an optimal price of $25, which is far below the actual national average price.

Figure 1
Van Westendorp PSM results for Cell 1 (Low price range, no outside reference)



We found that, for all four van Westendorp questions, the responses were influenced by the price range factor. Figure 2 displays box plots with groupings based on our reference price ranges (low, high, and no range) for one of the four PSM questions, "At what price would adult day care be getting expensive?" Comparing the low and high price ranges, we see that the median is greater for the high price range. Looking at the cells where no reference price range was forced, the responses are more dispersed, with some extreme outliers. However, the median prices for these cells are closer to those with the low price range than to those with the high price range.

Figure 2
Price responses for van Westendorp "getting expensive" question by experimental cell

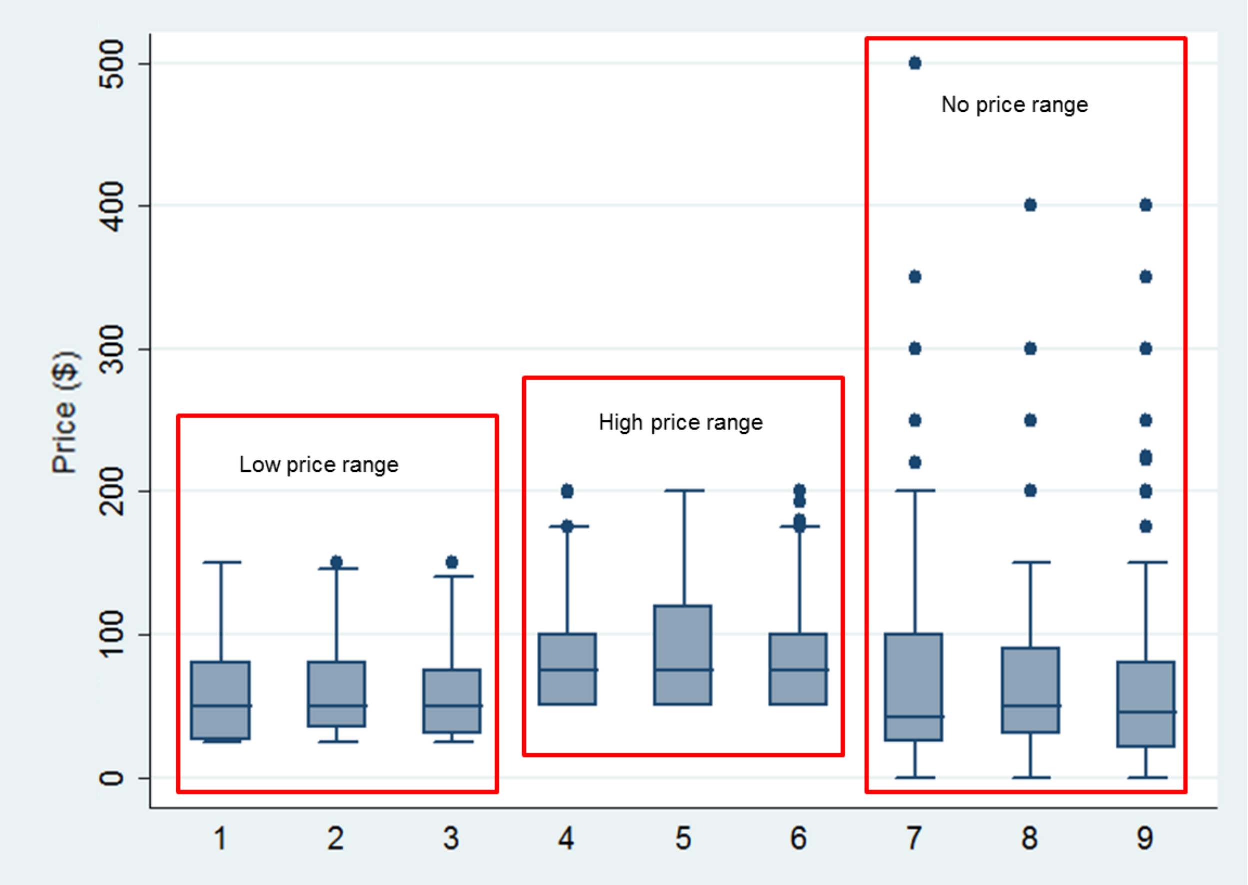
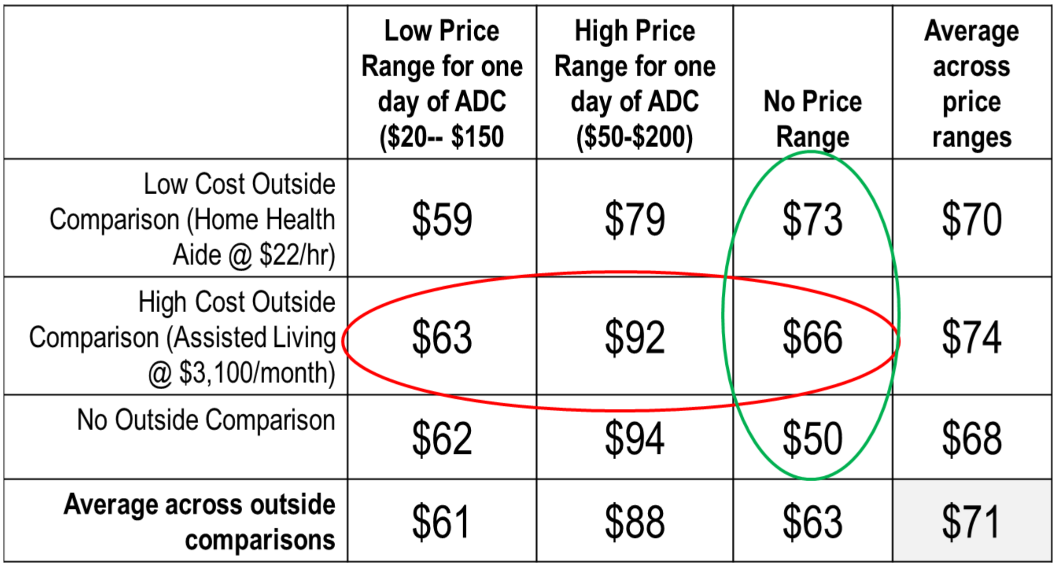


Figure 3 shows the mean price responses for the same "getting expensive" van Westendorp question across the experimental cells. Providing a higher price range for the van Westendorp question elicited higher prices. Results for the other three price questions are similar. While there was no main effect for the outside reference price variable, Figure 2 does reveal an interaction effect between this factor and the price range variable. When no price range information is present, the reference price information impacts the elicited willingness to pay. These effects were statistically significant at p<0.05. Additionally, under the high price range treatment, information about a low cost alternative (home health aide) appears to depress somewhat the willingness to pay for ADC.

Figure 3
Mean "Getting expensive" price response for price range and outside reference combinations



	Low Price Range for one day of ADC ($20-- $150	High Price Range for one day of ADC ($50-$200)	No Price Range	Average across price ranges
Low Cost Outside Comparison (Home Health Aide @ $22/hr)	$59	$79	$73	$70
High Cost Outside Comparison (Assisted Living @ $3,100/month)	$63	$92	$66	$74
No Outside Comparison	$62	$94	$50	$68
Average across outside comparisons	$61	$88	$63	$71

Experiment 2:

Effect of price range and reference price comparison on choice-based conjoint estimates of willingness to pay for adult day care services

The results from our first experiment were dramatic and we wondered if choice-based conjoint, a far more robust method for estimating willingness to pay, was subject to similar effects.

As noted previously, choice-based conjoint is an indirect method for estimating willingness to pay. Survey respondents are presented with a series of buying situations in which price is systematically varied; their choices are observed and incorporated into a statistical model of their decision processes. Prices are almost always presented as a series of point values over a specific range. For example, we might present respondents with prices of $25, $50, $75, $100, $125, and $150 for a single day of adult day care. This range encompasses the national average ($67) as well as the observed low ($27) and high ($150) prices. However, we might, for various business reasons, have decided that the lowest price of interest to us is $50, and we really want to know the potential for charging a much higher price, so we decide to set the prices at values between a low of $50 and a high of $200. Many of us (as we did) will assume that this difference will not impact our conjoint-based estimates of willingness to pay. If someone is only willing to pay $75 (and the competitive context remains the same), the values of the lowest and highest prices should not affect our model.

Price is typically one of a few (or perhaps several) variables incorporated into a conjoint exercise. One of the attractive features of choice-based conjoint is the ability to incorporate alternatives that do not share attributes or levels (usually in an "alternative-specific" design). For example, in a transportation choice, we might include several public transportation options (with different prices and other features, such as waiting times) as well as "constant" alternatives such as personal auto and bicycle. We assume that the choices a respondent makes are a function solely of the individual utilities of the alternatives. We decided to put this to the test by varying the availability of an "outside good." As in our first experiment, we used assisted living and home health aides as alternatives to adult day care.

We added one more element to the experiment. Considerable empirical evidence indicates that subtle differences in questionnaire content and flow can affect the answers that survey respondents give (for a summary, see Sudman, Bradburn and Schwarz, 1996). This may extend to the "pre-conditioning" information provided before a conjoint exercise. As long as all the subjects in our survey receive the same pre-conditioning information, we can be confident that any observed effects of our independent variables (price range and outside good) are not confounded with pre-conditioning information. However, if pre-conditioning interacts in any way with the other two-variables, we would be unable to detect such an effect. For that reason, we decided to include two variations of pre-conditioning, which we labeled "rich" and "poor."

We designed a 2 X 3 X 2 factorial experiment where, as in our first experiment, we varied the *range* of prices in the conjoint exercise and whether or not an *outside good* was included in the set of alternatives. Prices for adult day care in the conjoint exercise ranged from either $25 to $150 or from $50 to $200. With respect to the outside good, the levels were: no outside good (OG) — that is, a standard "none of the above" option; the addition of an assisted living (AL) option; or the addition of a home health aide (HHA) option. As noted above, we also varied the pre-conditioning information, with two levels reflecting differences in the amount of detail provided in the offer descriptions. For the "poor" information condition we described adult day care as follows:

> "**Adult day care services** provide health, social, and therapeutic activities in a supportive group environment. Services offered usually include social activities, meals and snacks, personal care, and recreational activities. Care can be obtained for one or more days per week, depending on your needs.

For the "rich" pre-conditioning, we added the following to the above description:

> "Adult day care services provide relief for caregivers, or may enable an aging individual to continue living at home if the primary caregiver has other responsibilities (e.g., full time job)."

We added similar sentences to the descriptions for assisted living and home health aide (outside good) in the "rich" pre-conditioning cells. In each case, the addition described a specific benefit, as in the example above.

We specified two hypotheses to test, following from the hypotheses tested in our first experiment:

> *H1:* Given that the price range in the conjoint design encompasses all relevant prices, the location of the range will have an impact on willingness to pay.
>
> *H2:* The presence of a comparison "outside good" will impact the probability of choosing an adult day care option at a given price point.

Within the overall experimental design, the conjoint exercise included these attributes: price (6 levels for ADC, 2 levels each for assisted living and home health aide); client-to-staff ratio for ADC; maximum number of clients for ADC; and whether or not transportation was included, available at extra cost, or not available (again, for ADC only). These additional ADC attributes were included so that we could vary the ADC offers and induce respondents to make trade-offs as they do in a typical conjoint study.

As with our first study, we embedded this experiment in one wave of our quarterly tracking study of approximately 1,500 U.S. adults. We used Sawtooth Software's CBC/HB module to estimate disaggregate models for each cell in the experimental design.

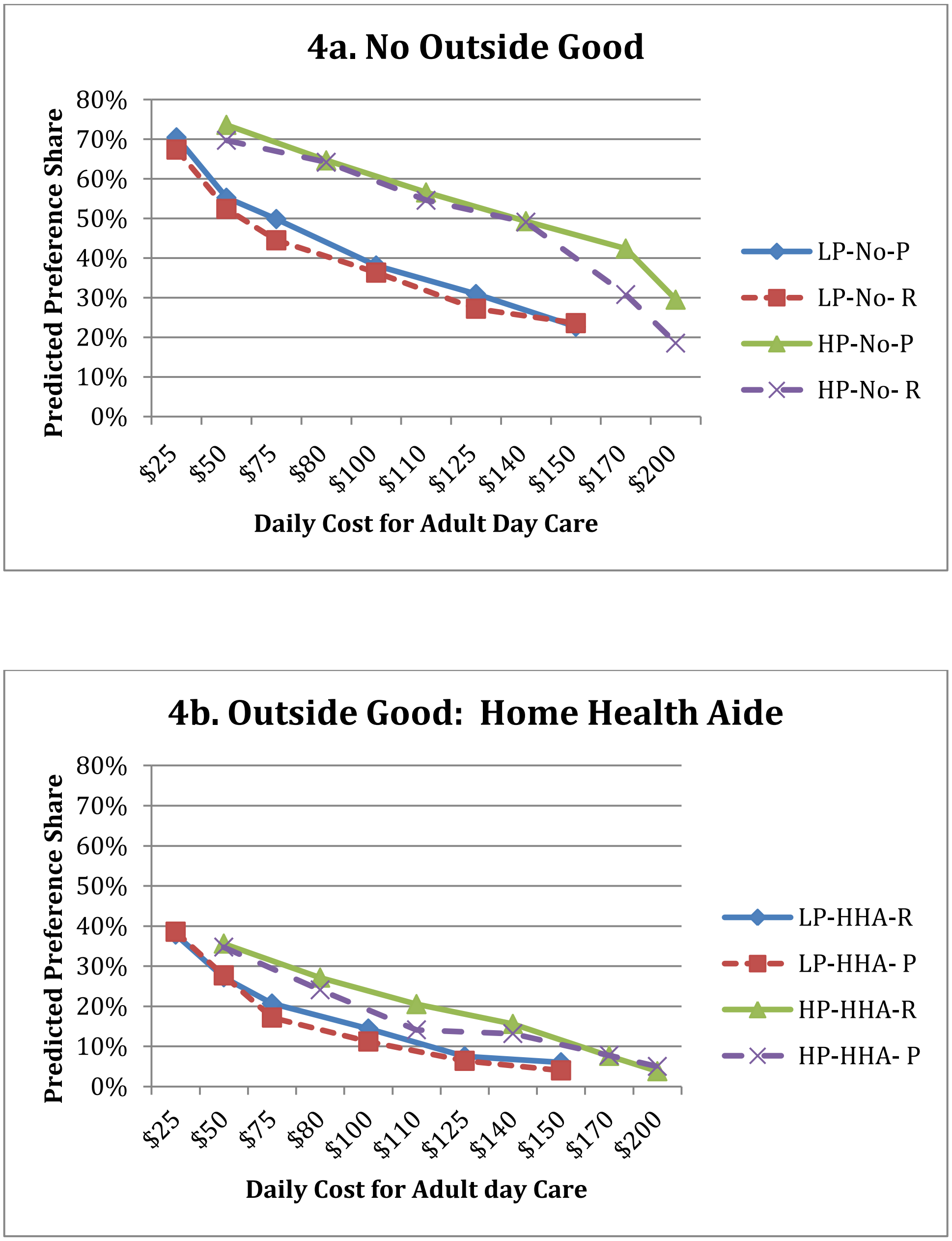
Finally, in order to provide a reference point to our first experiment, we asked a single van Westendorp question ("At what price would adult day care be getting expensive?) following the choice-based conjoint exercise. Respondents were randomly assigned to one of three conditions for this question: no price range; price range \$25-\$150, or price range \$50-\$200, as in our first experiment.

Experiment 2 Results

In order to identify the impact of the experimental variables, we used a market simulator to predict preference shares for a single adult day care alternative at varying prices against a "none of the above" option. Most marketers will make pricing decisions based on predicted or simulated preference or choice shares, and the actual part-worths cannot be compared across the experimental treatments because of differences in the scale parameter in the lower level (logit) model that estimates each individual's decision process.

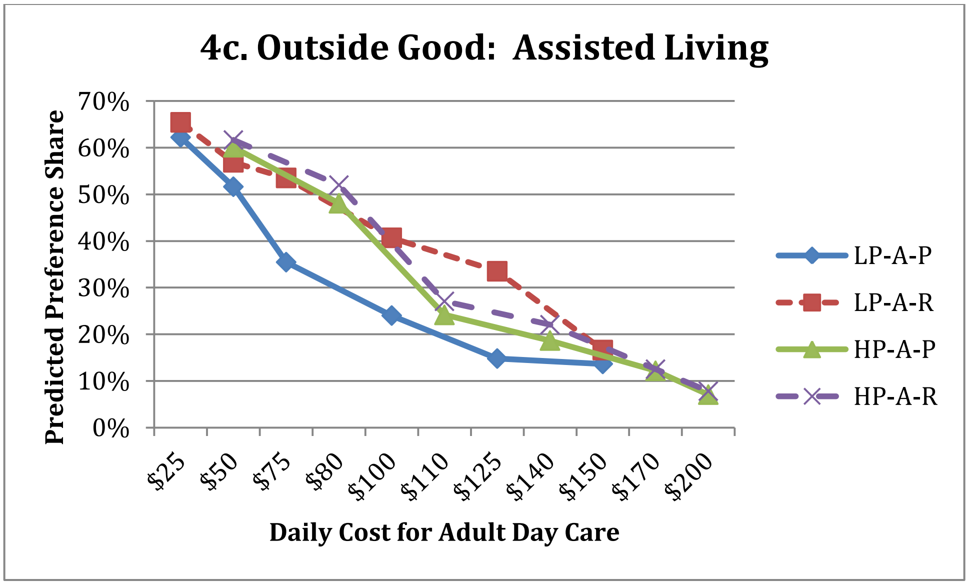
Results of these simulations are presented in Figures 4a through 4e. In these figures, the price range is indicated by "LP" (low price) and "HP" (high price). The type of outside good is indicated by "No" (no outside good), "HHA" (home health aide) and "A" (assisted living). Finally, the cells with rich and poor pre-conditioning information are indicated by "R" and "P" respectively. Price range has a dramatic impact on predicted preference shares when no outside good (apart from "none") was included in the choice task (Fig. 4a). Within the overlapping portion of the price ranges (i.e., \$50-\$150), respondents who saw the higher price range in their conjoint exercise are more likely to choose ADC over the "none" option at each price point. The effect is not so pronounced for the cells where the home health aide outside good was present (Fig. 4b) but we still see higher predicted preference shares for the cells with the high price range. The picture is not so clear when assisted living is offered as an alternative to adult day care (Fig. 4c). Specifically there appears to be an interaction effect such that "rich" preconditioning offsets the effect of the low price range. Overall, however, differences between cells with "rich" and "poor" pre-conditioning information are small.

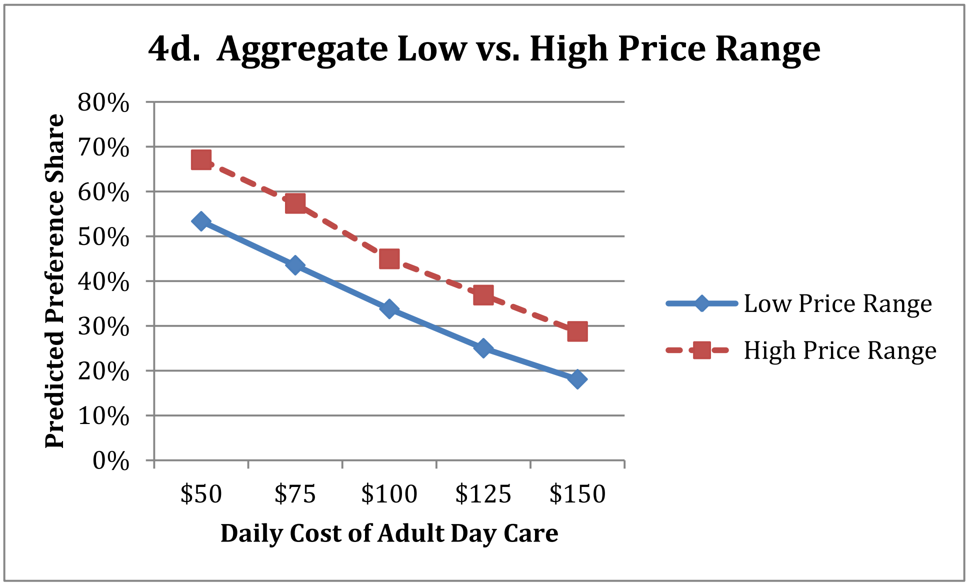
Figures 4a-e
Comparison of simulated preference shares across experimental treatments



4a. No Outside Good

4b. Outside Good: Home Health Aide


4c. Outside Good: Assisted Living
Predicted Preference Share
70%
60%
50%
40%
30%
20%
10%
0%
$25
$50
$75
$80
$100
$110
$125
$140
$150
$170
$200
Daily Cost for Adult Day Care
LP-A-P
LP-A-R
HP-A-P
HP-A-R


4d. Aggregate Low vs. High Price Range
Predicted Preference Share
80%
70%
60%
50%
40%
30%
20%
10%
0%
$50
$75
$100
$125
$150
Daily Cost of Adult Day Care
Low Price Range
High Price Range

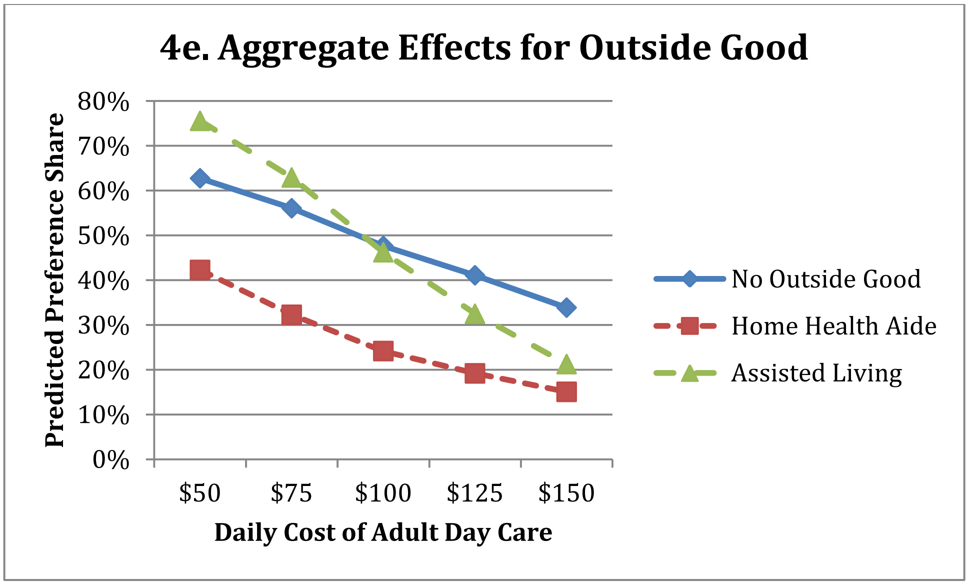


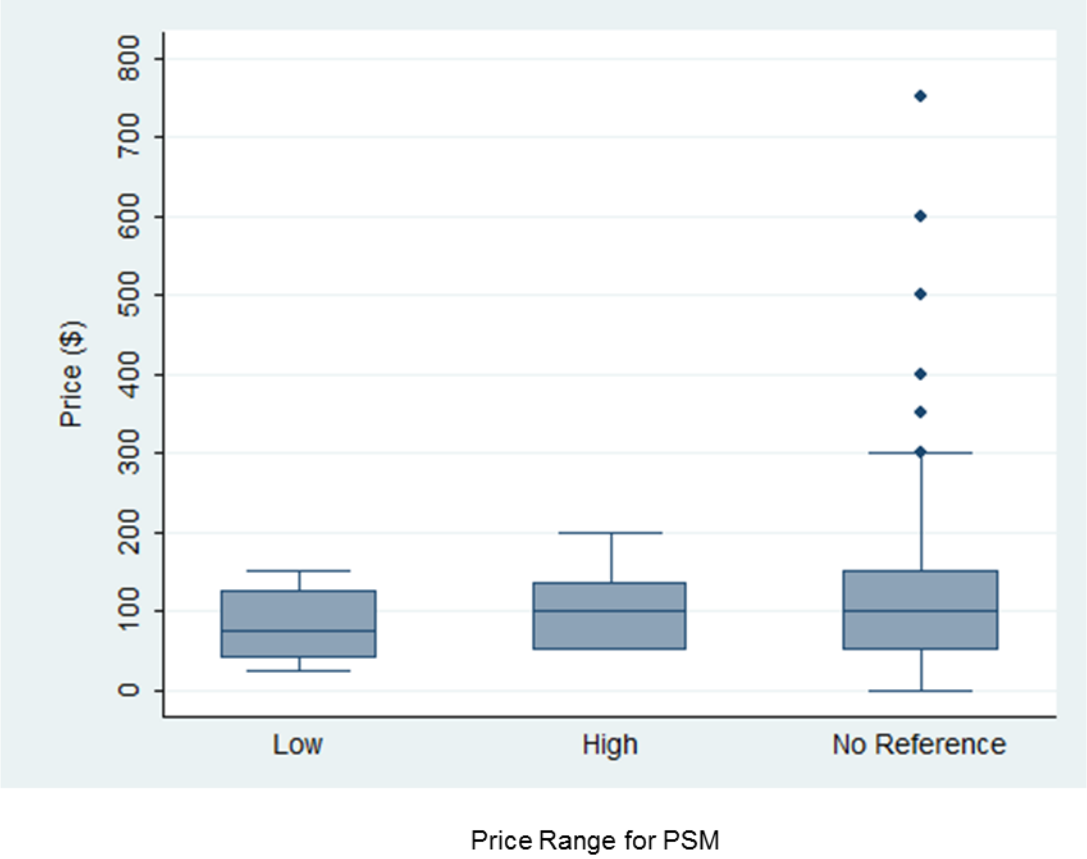
Figure 4d shows the *average* predicted preference share for all cells with the high price range compared to the average for all cells with the low price range (for the overlapping price points). These results suggest that the price range used in a choice-based conjoint exercise may systematically affect estimates of willingness to pay.

Figure 4e shows the average predicted preference share for each of the three levels of outside good. As with price range, the specification of the outside good appears to affect the overall probability of choosing an alternative. Moreover, there is an interaction. The slope of this curve for the assisted living cells is greater than for the other two treatments (which are essentially parallel to each other).

Because, for the experimental cells with an outside good, there were more concepts in each choice task (4 vs. 3), it is possible that differences in the scaling for "None" could lead to a spurious effect in those cells. While we cannot rule out this possibility, the results depicted in Figure 4e suggest that at least some of the effect is due to the presence of the specific outside good rather than the number of concepts in the task. If differences were due to the number of concepts, we would expect the average predicted shares for the home health aide and assisted living cells, with four concepts per task, to be similar and different from the predicted shares for no outside good (three concepts per task) instead of the pattern we see in these results.

Figure 5 displays the results from our single follow-up van Westendorp question. The median price for respondents who saw the higher price range in the PSM is higher than the median for respondents who saw the low price range, as in our first experiment. However, the median for respondents with no price range for reference is just about the same as for those who saw the high price range, in contrast to our first experiment where the median response for no price range was closer to the median for the low price range. The price range that respondents saw in the conjoint tasks did not have much impact on the responses to these questions, except when there was no price range for the PSM.

Figure 5
Post CBC-Exercise van Westendorp "Getting expensive" results



Discussion

Frankly, we find the results of our experiments disturbing. While we expected the van Westendorp PSM to be sensitive to the location of a constrained price range, we believed that choice-based conjoint would be much less susceptible to such extraneous influences.

Irrational Decisions and the Van Westendorp PSM

With respect to the PSM, our experiment reinforces our skepticism about the validity of this method for determining the price to charge for a product or service. For products that are entirely new or for which there are no near substitutes or other price references, PSM results are often well off the mark, at least in terms of marketers' *a priori* assumptions about what they can charge. The fact that the *average* elicited price when we provided a slightly higher reference range for the PSM is more than 40% higher than the average elicited with a lower reference range (or with no reference range at all) is cause for concern.

Irrational Decisions and Choice-Based Conjoint Pricing Models

The theoretical foundation for the discrete choice model holds that the probability of choosing an alternative is a function of the utility of that alternative relative to the utilities of the other alternatives (specifically, this likelihood is a function of the ratios of the antilogs of these utilities). Therefore we would not expect different price ranges to lead to different choice probabilities as long as those ranges incorporate the relevant set of prices. All other things being equal, we would expect the probability of choosing adult day care when the price is $75 to be the same whether the range of tested prices is $25 to $150 or $50 to $200. However, our experiment shows that choice-based conjoint is susceptible to seemingly extraneous factors. One possible explanation for this finding is that, in the absence of knowledge about the price of adult day care, the respondents infer that the range of prices presented represents a distribution around some average price.

The impact of varying an "outside good" in the choice model is somewhat perplexing. In order to compare the different treatments in our second experiment we ran simulations to predict preference share with only two alternatives: an adult day care option (with varying prices) and a "none" option because we had different models for each of the cells (i.e., some had part-worths for an assisted living option, some had part-worths for the home health aide alternative). The observed impact of the outside good may reflect price "anchoring." Adult day care seems like a better deal at any price point when compared to the higher-priced assisted living, but not such a good deal when presented alongside the lower-priced home health aide. Alternatively, as might be the case in our first experiment with the van Westendorp PSM, the respondents could be using the outside good to determine how much adult day care should cost.

A common practice in choice modeling is to identify the *realistic* lower and upper bounds of the price range and then to increase that range somewhat by extending both the lower and upper boundaries. The results of our experiment indicate that introducing any asymmetry in these extensions (that is extending by a greater percentage at one end than at the other end) could have an impact on estimates of willingness to pay derived from the choice model.

IMPLICATIONS AND RECOMMENDATIONS

Because pricing is critical to every company's success, the guidance provided by research must lead marketers to make the best possible bets in setting prices. Our experiment with the van Westendorp PSM leads us to conclude that this method is unpredictable in that extraneous factors influence the elicited prices. There has been little systematic research into the conditions that might lead to either more or less reliable and valid pricing information from this technique. Without more research, we cannot define any "best practices" that will improve the validity of this method. That being said, for situations where more robust methods are impractical, *direct elicitation* of willingness to pay may provide some guidance for developing a pricing strategy. For example, asking either or both van Westendorp questions around "expensiveness" may provide an estimate of the distribution of reservation prices. In such cases we recommend that researchers add a purchase likelihood measure tied to the respondent's elicited price, and that the questions ask for willingness to pay for the *value proposition* or *delivered benefits* rather than for the *product.* For the PSM results we obtained, using the points of marginal cheapness, expensiveness, optimum price, and indifference likely would lead to a bad pricing decision, so asking all four van Westendorp questions is probably unnecessary.

Our second experiment reinforces our belief that choice-based pricing models need to incorporate as much realism as possible. Thus, when there are competitive alternatives that might lie outside the target product category (such as assisted living, in this case), we should try to account for those alternatives explicitly, perhaps with extended alternative-specific designs or estimation of nested models. In setting price ranges to test in a study, knowing that a shift in the location of the range might impact the results should lead us to take more care in determining just what those ranges should be.

Choice-based pricing models are fundamentally different from direct elicitation methods. If we find a potential bias, we have the flexibility to change our model specification or estimation procedure to improve the model. Choice-modeling is the subject of hundreds of academic studies, and the underlying mechanics of choice models allow us to test different theories of consumer-decision making. For example, in many situations consumers may not follow the *compensatory* choice rule that is the foundation of the utility maximizing process assumed by the most commonly applied discrete choice models. In response, researchers have developed new models that can account for both compensatory and non-compensatory choice processes (notably, Gilbride and Allenby, 2004).

Our results also suggest that respondents are very sensitive—consciously or not—to all the information presented in the survey. For example, the willingness to pay estimates we observed in the van Westendorp question that followed the conjoint exercise indicate that the price range information provided in that question served to "reset" the respondents in some way. Thus, we found that the price range information provided in that question had more influence on their answers than did the price information (in the form of low or high ranges) they encountered in the twelve conjoint tasks. However, if no price range was provided in this question, the conjoint price range hand a small effect on the average "getting expensive" price.

The ultimate goal of pricing research is to guide marketers to better pricing decisions. Market research methods provide an estimate of the amount that customers are willing to pay, but that estimate is only one piece of the pricing puzzle. We would hope that a marketer looking at our van Westendorp results would take her prior beliefs (i.e., the current market pricing) into account and, in this case reject the Optimum Price Point suggested by the PSM.

Our confidence in our methods is enhanced by an understanding of the potential biases of the method. Knowing the magnitude and direction of these biases allows us to compensate for them in some way, or modify our methods to reduce or eliminate their impact. Potential biases and errors in choice-based conjoint have been the subject, as we observed, of much academic research. We cannot say the same for the van Westendorp PSM.

References

Ariely, D. (2008). *Predictably Irrational: The Hidden Forces that Shape Our Decisions.* New York: Harper.

Gilbride, T.J. and Allenby, G.M. (2004) A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, **23**, 391-406.

MetLife Mature Market Institute (2009), *The 2009 MetLife Market Survey of Nursing Home, Assisted Living, Adult Day Services and Home Care Costs.* www.MatureMarketInstitute.com.

Miller, K.M., Hofstetter, R., Krohmer, H. and Zhang, J. (2011) How should consumers' willingness to pay be measured? An empirical comparison of state-of-the-art approaches. *Journal of Marketing Research*, **48**, 172-184.

Ormerod, P. (2005). *Why Most Things Fail.* New York: John Wiley & Sons.

Sudman, S., Bradburn, N.M., and Schwarz, N. (1996) *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology.* Sand Francisco: Jossey-Bass Publishers.

OPTIMIZING PRICING OF MOBILE APPS WITH MULTIPLE THRESHOLDS IN ANCHORED MAXDIFF

EDWARD PAUL JOHNSON
SURVEY SAMPLING INTERNATIONAL
BRENT FULLER
THE MODELLERS

INTRODUCTION TO ANCHORED MAXIMUM DIFFERENCE

Maximum Difference models were proposed and practiced starting with Jordan Louviere in the early 1990s (Louviere 1991; Louviere 1993). Since that time they have been growing in popularity and are still common practice today because they have the advantages of discrete choice models: avoiding scale response bias and forcing respondents to make trade-offs. However, one criticism that emerged of the model is that the resulting utilities were only relative. People could have a high utility for items that they hate because they are relatively better than the other alternatives that they hate even more (Lattery, 2010). While this may not be terribly problematic for segmentation purposes, it is not useful for finding how many people actually like the item. In the past few years a few methods of anchoring the utilities to a threshold have been suggested:

1) *Including a Threshold as a Design Item* – this approach was suggested by Bryan Orme where a set dollar amount or statement of no change is included in the set of items shown to a respondent (Sawtooth Software, 2006; Orme, 2003).

2) *Dual Response (Indirect) Approach* – this approach was recommended by Jordan Louviere (as a technique he has long used in his practice) in a conversation with Bryan Orme and resulted in a paper that suggested asking a question about the threshold after every task (Orme, 2009).

3) *Direct Binary Approach* – this approach was suggested by Kevin Lattery where the respondents were asked about the threshold in relation to each item at the end (Lattery, 2011).

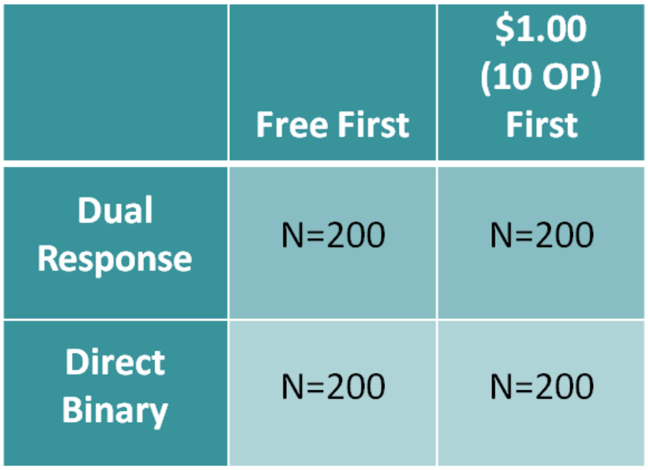
Each of these methods identifies which items in the design space meet the threshold by comparing each respondent's individual utilities to a threshold. As a result, if a respondent does not like any of the items in the design it can be identified because the threshold utility will be the highest. The utilities can then be used to do a TURF analysis. This technique can then answer some questions depending on the type of threshold: how many people consider each item important (preference threshold), what percentage of respondents would spend $5 to add a certain benefit to their policy (pricing threshold), and which issues would be important enough to get people to vote early (action threshold). In this particular example we used pricing thresholds for a mobile application for survey taking.

Introduction to Mobile Surveys

Mobile devices continue to rapidly expand globally. Nielsen estimates that over half of the US adult population who has a mobile phone owns a smartphone and that percentage is on the rise (Nielsen, 2012). As a result more and more research companies are trying to engage panelists in survey opportunities via a mobile device. Survey Sampling International (SSI) has a history of implementing mobile technology in survey taking opportunities including standard panel research (Johnson 2011), passive tracking technology (Johnson and Wright 2011) and occasion-based research (Hadlock et al, 2012). However, SSI had not launched a mobile app for its online US panels yet, in particular Opinion Outpost. As the iPhone mobile app was developed, SSI wanted to know if panelists should be charged for downloading the app and how the survey app compared to other popular apps among Opinion Outpost iPhone users. Thus, SSI decided to do a Maximum Difference exercise on various mobile applications including the new Opinion Outpost app. However, instead of trying to implement a free and a paid level by including them as items in the design, we wanted to try to implement an extension of the Direct Binary and the Dual Response approaches by adding a second threshold level.

Methods

The survey was fielded in October of 2011 with 800 online panelists from SSI's North American panel Opinion Outpost. We wanted to test both the Direct Binary approach and the Dual Response approach to adding additional threshold levels. Because a previous experiment showed significant differences between the two methods, we wanted to see if the difference could also have been caused by an order effect. So half of the panelists in each method were shown the free download threshold first while the other half were shown the download for $1.00 (10 Opinion Points) first. The resulting design with sample sizes is shown in Figure 1.



	Free First	$1.00 (10 OP) First
Dual Response	N=200	N=200
Direct Binary	N=200	N=200

Figure 1. Sample Size and Experimental Design of the Study

All four experimental groups received the same 16 mobile apps (shown in Table 1) on 12 screens with four apps shown on each screen. We only showed four items at a time when comparing the items directly to each threshold to minimize any effect from the number of items shown in the direct binary method. Also, at Kevin Lattery's suggestion, we implemented constraints on which of the items were shown against the second threshold to avoid logical inconsistencies. If the download for free threshold was shown first, then any item not selected would be auto coded as also below the second threshold (download for $1.00) and not shown to the respondent. Likewise, if the download for $1.00 threshold was shown first, then any item selected would auto coded as above the second threshold (download for free) and not shown to the respondent. This implementation should have lowered the respondent burden as well as forced logical consistency in the thresholds.

Apps Tested
Opinion Outpost
LinkedIn
20 Minute Meals - Jamie Oliver
Angry Birds
BillTracker
Yahoo! Finance
CNN App for iPhone (U.S.)
Surf Report
Lose It!
Urbanspoon
Entertainment Tonight - ET
SCRABBLE
Plants vs. Zombies
Convert - the unit calculator
New Oxford American Dictionary
FedEx Mobile for iPhone

Table 1. Applications Tested in Experiment

After the data was collected the utilities were estimated using the HB model as described by Kevin Lattery (Lattery, 2010). The utilities for each method were then compared by creating a scatterplot of the average utilities for each item by method. Also the percent of the sample that rated the app above each threshold was also used to compare the methods.

Comparison of Average Utilities

First we looked to see if the utilities calculated for the mobile apps were consistent across the four groups. Figure 2 shows the average utility of each app for the four experimental groups. The first item of note is that all the correlations are close to 1. Because they are all close to one the methods predict each other well and seem to be doing the same thing. It is also interesting to note that the scale of the utilities does not seem to change by the order effect (the slopes are close to 1), but the direct binary method tends to have more differentiation in the utilities resulting in a slope of .9 against the indirect (dual response) method. This makes sense as there would tend to be less noise when a respondent is asked a question directly with logical constraints than when they answer indirectly. Still, overall all the methods seem to be comparable.

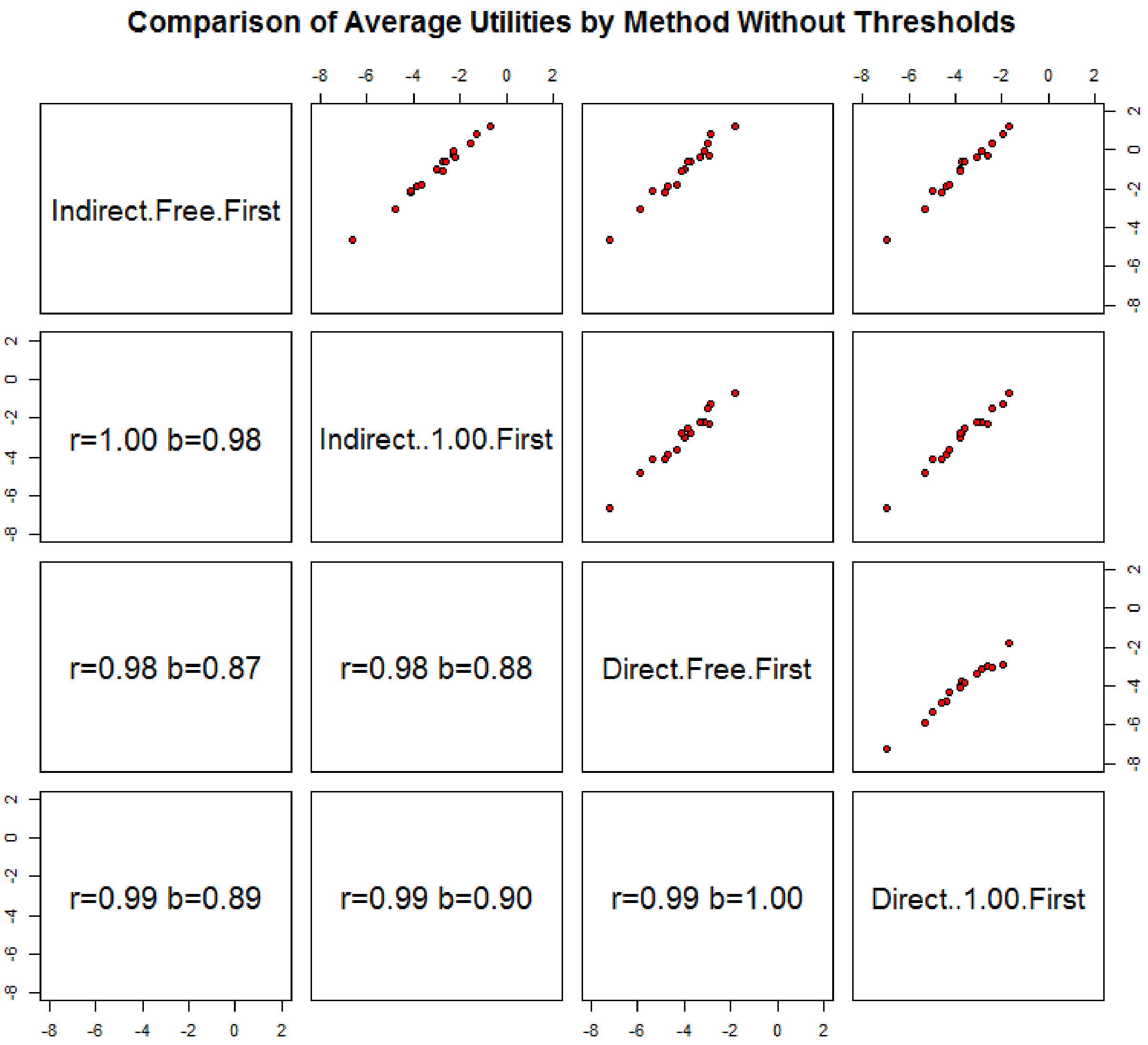


Figure 2. Average Utilities of Mobile Apps by Experimental Group

However, when the thresholds are added into the model the results for the experimental groups differ significantly. Figure 3 shows how most of the correlations decrease when the thresholds are added. The noticeable exceptions are the two Direct Binary groups. Here the slope and the correlations are both very close to 1, indicating there are no order effects for the Direct Binary approach. However, there seem to be significant order effects in the indirect method. In particular the respondents that saw the $1.00 threshold first seemed to not distinguish at all between the $1.00 threshold and the free threshold. While the respondents that saw the free threshold first were more likely to distinguish between the thresholds, they did not make as large of a distinction as the direct binary respondents.

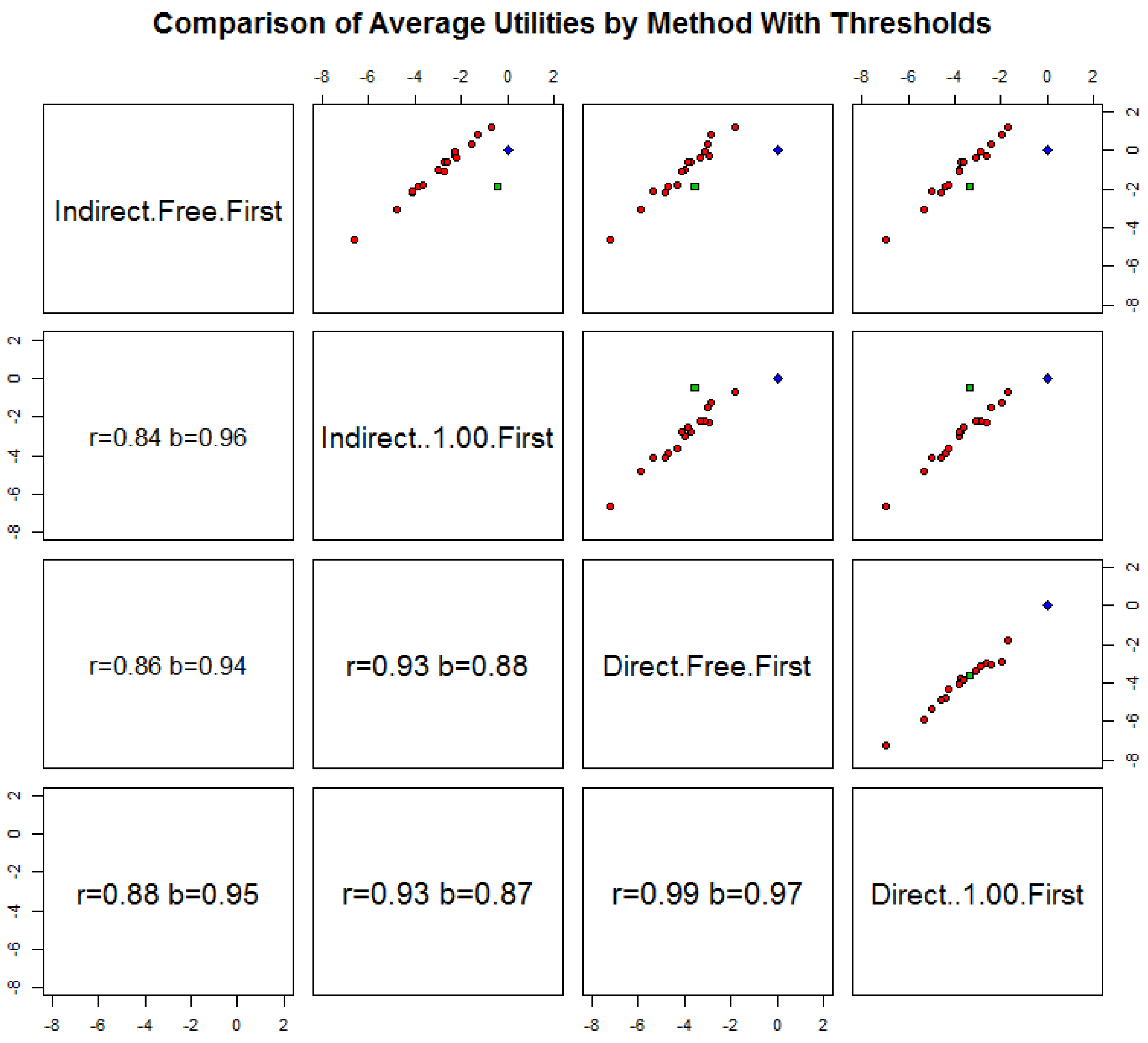


Figure 3. Average Utilities of Mobile Apps and Thresholds by Experimental Group

Comparison of Download Results

Figures 4 and 5 show how the take rates in each method translate into actual decisions. The take rates in all four groups directionally follow each other. The most popular app (Opinion Outpost) is most popular for all four groups, followed by well-known gaming apps (Angry Birds and Scrabble) with a surfing app at the bottom (Surf Report). Furthermore, there seems to be no order effect in the Direct Binary group as the free take rates are similar. However, there are order effects in the Dual Response groups because the take rates in these two groups do not agree. In particular the group that saw the app for free first actually reflects a higher percentage willing to download for free while the group that saw the $1.00 threshold first is projected to have a much lower probability of downloading for free.

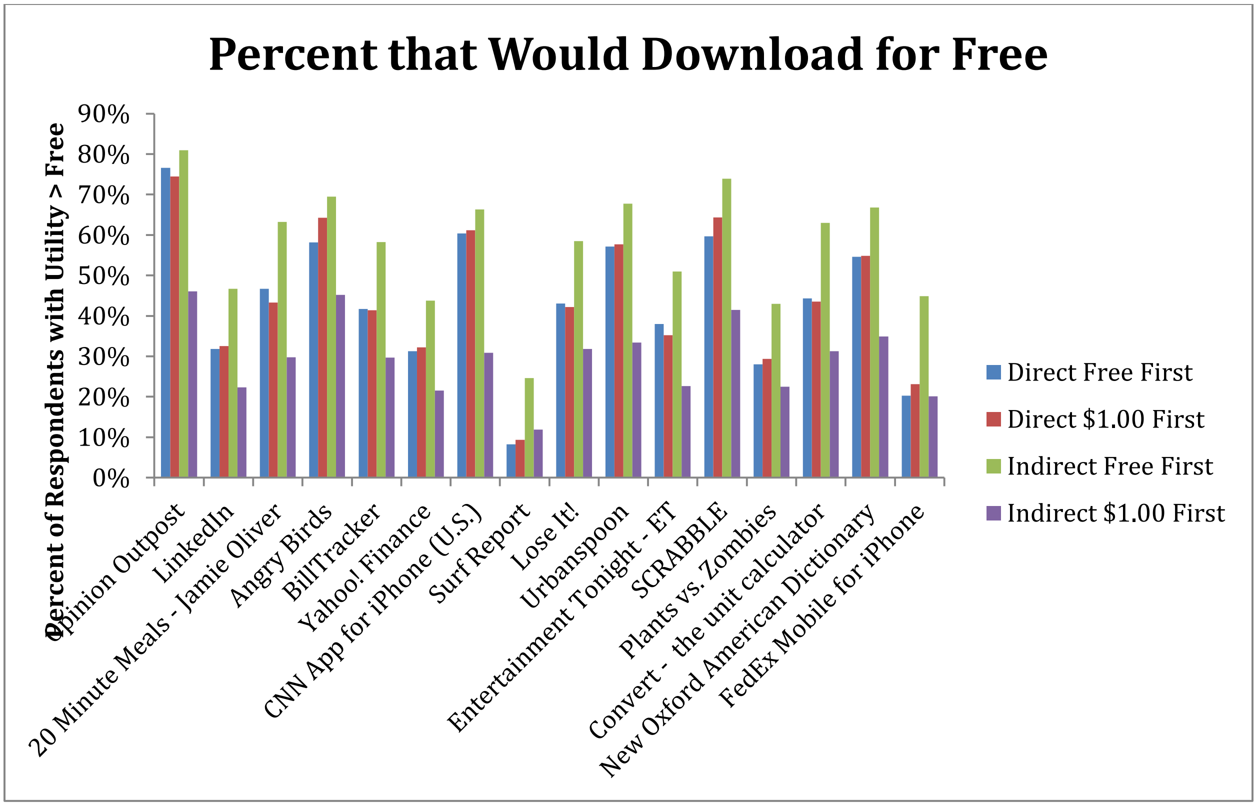


Figure 4. Percent Willing to Download Each Mobile App for Free

The Dual Response method of elicitation led to results that were extremely price insensitive. Figure 5 show how the percentage of respondents who would be willing to download the app for $1.00 is much higher in the Dual Response groups than in the Direct Binary groups. This switch is because respondents who received the Dual Response method of elicitation are represented as not price sensitive to the app. Based on the Dual Response data, SSI would have concluded that they should charge for the app and collect the $1.00 per download to offset development costs because the panelists didn't really care about $1.00 (take rate dropped 10% in the free first group and only 2% in the $1.00 first group). The Direct Binary groups show that the interest in the app falls dramatically when the respondent actually has to pay anything for the app. With this data SSI would have concluded to not charge for the app because it the 50% drop in the take rate wouldn't be worth the extra dollar they would have collected by charging for the app. These method effects are very significant to the point that the business decision rule would have changed.

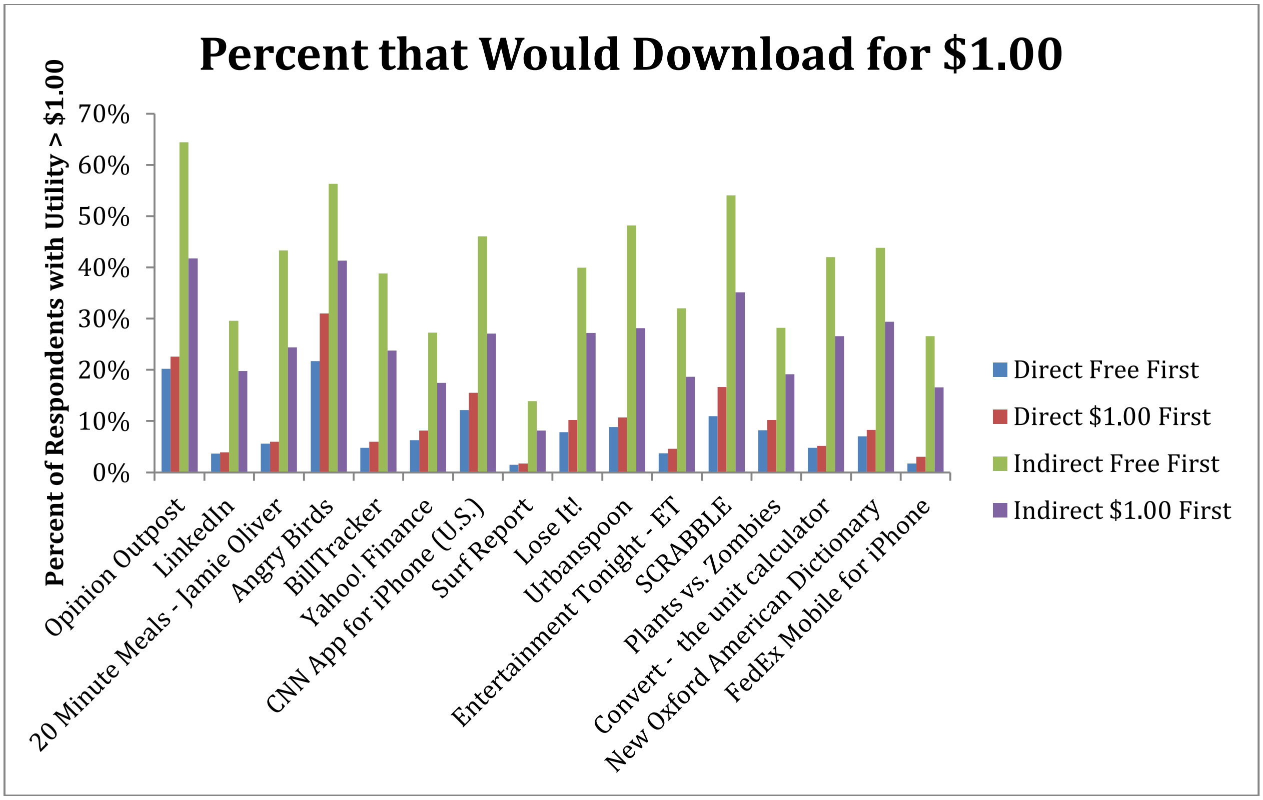


Figure 5. Percent Willing to Download Each Mobile App for $1.00

To summarize, there are minimal order effects in the Direct Binary method, but the order effects in the Dual Response method are very significant. In particular the Dual Response respondents that saw the $1.00 threshold first were much less likely to be projected to actually download any app for free or for a dollar. These order effects lead us to have lower confidence in the data gathered through the Dual Response method. Potential reasons for this order effect will be discuss in the next section.

On the positive note for SSI, the interest in the Opinion Outpost app remained high across all the methods and competed well against other apps for panelist time. It promotes the idea that Opinion Outpost panelists who have a smartphone want to access surveys through a mobile app and will spend time taking surveys on their mobile device.

Discussion and Conclusions

In this study we were able to show that both methods of collecting data actually produced similar relative utilities for the individual apps. In general the Direct Binary method had app utilities that were more stretched out with less noise than the Dual Response app utilities. These are consistent with the findings from Kevin Lattery's comparison. However, the thresholds in this study differed significantly based on method. In particular the Dual Response method resulted in respondents seeming less price sensitive to the point that the business decision rule would have changed. The significant order effects in the Dual Response method lead us to believe that the Direct Binary captures the correct price sensitivity, so SSI decided to not charge for the app when it launches.

While the utilities of the thresholds differed, it could be caused, not by the analytical methodology, but rather by the presentation method to the respondent. For example, the panelists might not have noticed that the threshold wording changed half way through the exercise. This example would explain why the respondents are not differentiating between the two thresholds in the Dual Response method, leading to price insensitivity. While we cannot prove that this theory is true because we didn't ask any open ends about the maximum difference section, it is important to look at in future research. While it is easy to blame the panelists for not noticing the threshold change, it is much harder to think of ways to change the survey design to meet the respondents' needs. In this case we think we could have made the difference clearer by breaking up the two sections (one for each threshold) of the Maximum Difference with a few questions, emphasizing the new threshold through bolding and bigger font size, or rewording instructions to make them clearer. We believe that these changes could make the Dual Response method fall more in line with the Direct Binary method even if there are two thresholds rather than just one.

Another more philosophical issue was raised by using two thresholds instead of just one. One possible concern is that including a second threshold interferes with the first threshold. Others would argue that you always have an anchor (just making the anchor a threshold level makes it easier to use) and that adding another threshold will not change the preference of the other items in relationship to each other. This philosophy would argue that the only difference in the three methods of anchoring the utilities to get a take rate would be in the presentation to respondents. In this case the method that should be used would be one that makes the most sense to respondents and is easier for them to follow. For example if the threshold is an importance threshold it might be awkward for respondents to see the words "is important" included in a list of potential features for a product, in which case either the Dual Response or Direct Binary approach might be more useful. Likewise if respondents are not used to the threshold changing in the middle of a maximum difference exercise the Direct Binary approach might be more useful than the Dual Response approach. Questionnaire pre-testing with open ends to elicit the respondents' thought processes can be effective in discovering the right technique to use in any certain survey. In future studies we may compare the scaling effect of adding the additional baseline either nested inside the design or asked externally in either method. It would also be interesting to note if changing the anchor to even be one of the non-threshold items would actual change the scale or the results of a TURF analysis. In either case more work should be done to consider exactly the effect that adding the anchor does to the Maximum Difference model.

References

Hadlock, Weston, Johnson, Paul, Roberts, Meghann, and Shea, Carol (2012), "Matching Data Collection Method to Purpose: In the Moment Data Collection with Mobile Devices for Occasioned Based Analysis" *2012 American Association of Public Opinion Researchers Annual Conference*, Orlando, Florida.

Johnson, Paul (2011), "Application of Mobile Survey Research for the Entertainment Industry" *Alert! Magazine*, Vol. 51 (4).

Johnson, Paul & Wright, Miles (2011), "Geo-Location Triggered Survey Research: Using Mobile Panelists' GPS Data to Trigger Survey Opportunities" *2011 CASRO Technology Conference*, New York City, New York.

Lattery, Kevin (2010), "Anchoring Maximum Difference Scaling Against a Threshold – Dual Response and Direct Binary Responses" *2010 Sawtooth Software Conference Proceedings*, 91-106.

Louviere, Jordan (1991), "Best-Worst Scaling: A Model for the Largest Difference Judgments" Working Paper, University of Alberta.

Louviere, Jordan (1993), "The Best-Worst or Maximum Difference Measurement Model: Applications to Behavioral Research in Marketing" *The American Marketing Association's 1993 Behavioral Research Conference*, Phoenix, Arizona.

Nielsen (2012), "America's New Mobile Majority: a Look at Smartphone Owners in the U.S." *Nielsenwire*, blog post on 5/7/2012 and can be retrieved at http://blog.nielsen.com/nielsenwire/online_mobile/who-owns-smartphones-in-the-us/.

Orme, Bryan (2009), "Anchored Scaling in MaxDiff Using Dual Response" *Sawtooth Software Research Paper Series*.

Orme, Bryan (2003), "Scaling Multiple Items: Monadic Ratings vs. Paired Comparisons" *2003 Sawtooth Software Conference Proceedings*, 43-60.

Sawtooth Software (2006), "SSI Web User Manual v5.0" Sequim, WA.

CONTINUED INVESTIGATION INTO THE ROLE OF THE "ANCHOR" IN MAXDIFF AND RELATED TRADEOFF EXERCISES

JACK HORNE,
BOB RAYNER,
REG BAKER, AND SILVO LENART
MARKET STRATEGIES INTERNATIONAL

INTRODUCTION

Maximum-difference scaling (MaxDiff) or Best-Worst scaling has been an effective choice-based method for uncovering the relative value among a group of items since Jordan Louviere first introduced it more than 20 years ago (Louviere, 1991). The technique appears to remove scale-use bias and does a better job of differentiating the relative worth of a group of items than rating scale-based methods. Since its introduction it has been used not only in determining relative worths, but in segmentation (Cohen, 2003) and in pricing (Chrzan, 2004; Johnson and Fuller, 2012).

Beginning late in the last decade, researchers began to question whether it was possible to "anchor" these relative worths to a fixed reference point. Were all items important or of value; were all items unimportant; or were some items important while others were not? The first methods developed sought to answer these questions by fusing rating scale data with MaxDiff results (Bacon, *et al*, 2007; Magidson, *et al*, 2009). Jordan Louviere, however, proposed a method at the 2009 Sawtooth Software Conference that used a dual-response question to anchor or calibrate MaxDiff results (Orme, 2009b). This became perhaps the first widely-used anchoring method for these kinds of data. More recently, Kevin Lattery introduced another more direct anchoring method which asked respondents to check whether each item is above or below some self-defined threshold (Lattery, 2010). This method has become known as "Direct Binary" in contrast to Louviere's "Dual Response" (Figure 1).

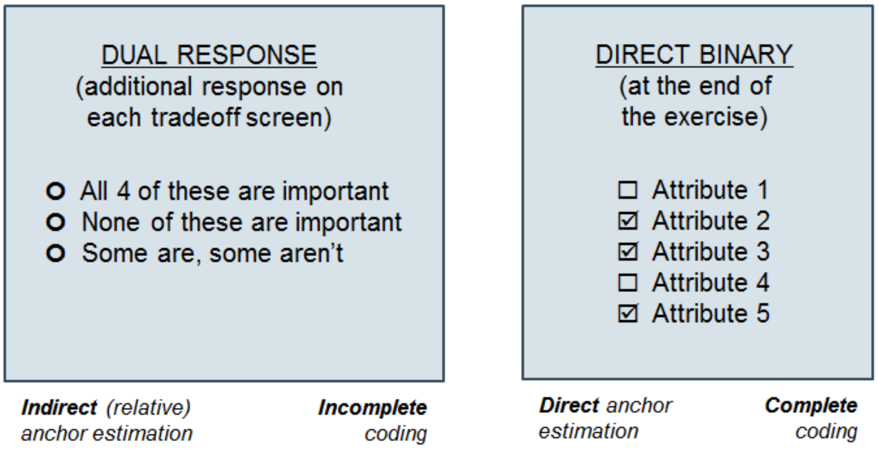


Figure 1. Dual response and direct binary exercises used to anchor tradeoff responses.

Our aim in the current paper is to extend the investigation into the efficacy of these anchoring methods. In so doing, we seek to answer the following questions:

- What happens if we dispense with tradeoffs entirely and only use the direct binary question to gauge relative rankings and the anchor/threshold?
- What is the effect of the number of items on each tradeoff screen in estimating the anchor/threshold?
- Are there alternative coding methods that can be used with the existing anchored tradeoff methods?
- Are there other, as yet unexplored tradeoff methods, that result in both relative rankings and an anchor/threshold?
- Is a cultural or scale-bias effect re-introduced as a result of the specific anchoring method used?

We examine actual respondent data along with simulated data in order to answer these questions. Ultimately, we seek to form a better understanding about which tradeoff and anchoring methods we ought to be using, and under what circumstances.

Experiment 1: Methods Comparison Using Actual Survey Data

Data Collection and Experimental Design

A split-sample experiment was used to test six different choice-based tradeoff techniques. A total of 1,808 respondents completed the survey, with approximately 300 respondents participating in each cell. Data collection methods included the following:

- MaxDiff (4 items per task); Dual Response Anchor
- MaxDiff (4 items per task); Direct Binary Anchor
- Direct Binary Only (no preceding tradeoff screens)
- Paired Comparisons (2 items per task); Dual Response Anchor
- Paired Comparisons (2 items per task); Direct Binary Anchor
- Two-By-Four (2 items per task); A paired comparisons method used by the authors that results in relative ranks and anchoring from a single response.

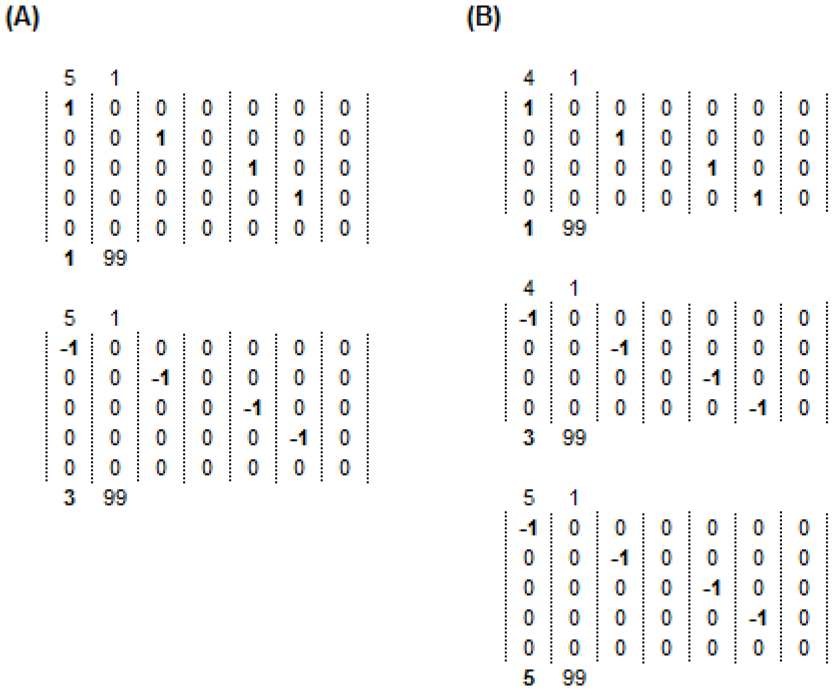
For each of these methods, respondents evaluated the same group of 17 attributes. Attributes involved personalized content, such as news, weather, sports, shopping, etc., that could be added to a web application such as email.

The MaxDiff methods used a design where 10 tasks were evaluated by each respondent, with four items shown in each task. The paired comparisons methods, including Two-by-Four, used a design where 20 tasks were evaluated by each respondent, with two items shown in each task. In all these conditions, respondents evaluated each item in tradeoffs an average of 2.35 times. The direct binary only condition did not require an underlying design, and respondents evaluated each item only once in a single grid question.

While there were six experimental conditions, we used two coding methods (see next section for details) for each of the three methods involving a direct binary anchor. This brought the total number of methods under investigation in this experiment to nine. All methods were analyzed using Hierarchical Bayes (HB) with Sawtooth Software CBC/HB, and using aggregate logit modeling.

Coding Methods

MaxDiff with Dual-Response Anchoring was coded as described in Orme (2009b). Imagine a situation where seven attributes are under consideration and the respondent evaluates four of these (attributes 1, 3, 5 and 6) in a given choice task. The respondent chooses attribute 1 as “best,” attribute 5 as “worst” and selects “some are important, some are not” in the dual response question. This set of choices is coded as in Figure 2A. The first choice set in this example indicates that the respondent selected attribute 1 over the other three attributes shown and over the reference or anchor row (vector of zeroes at the bottom). The second choice set indicates that the respondent selected attribute 5 as being *worse than* the other three attributes and the reference. This method is sometimes described as having incomplete coding because while we know where the items chosen as “best” and “worst” are relative to each other, relative to the other attributes and relative to the anchor, we do not know from either the responses or the coding where the two items not chosen are relative to each other or relative to the anchor.



(A)

```
5   1
1   0   0   0   0   0   0
0   0   1   0   0   0   0
0   0   0   0   1   0   0
0   0   0   0   0   1   0
0   0   0   0   0   0   0
1   99

5   1
-1  0   0   0   0   0   0
0   0   -1  0   0   0   0
0   0   0   0   -1  0   0
0   0   0   0   0   -1  0
0   0   0   0   0   0   0
3   99
```

(B)

```
4   1
1   0   0   0   0   0   0
0   0   1   0   0   0   0
0   0   0   0   1   0   0
0   0   0   0   0   1   0
1   99

4   1
-1  0   0   0   0   0   0
0   0   -1  0   0   0   0
0   0   0   0   -1  0   0
0   0   0   0   0   -1  0
3   99

5   1
-1  0   0   0   0   0   0
0   0   -1  0   0   0   0
0   0   0   0   -1  0   0
0   0   0   0   0   -1  0
0   0   0   0   0   0   0
5   99
```

Figure 2. *.cho file coding for MaxDiff with dual response anchoring; 7 attributes in all; attributes 1, 3, 5 and 6 shown; attribute 1 chosen as "best," attribute 5 chosen as "worst;" **Panel A**: "some are important, some are not;" **Panel B**: "all 4 are important."

Now consider the same choice situation, with the same "best" and "worst" responses, but the respondent selects "all 4 of these are important" in the dual response question. This set of choices is coded as in Figure 2B. Notice that the reference row has been removed from the first two choice sets. The selection of attribute 1 in the first set now indicates that attribute to be better than the other three. But, it no longer indicates that attribute, and that attribute alone, to be better than the reference. Likewise for attribute 5 in the second choice set. It is now worse than the other three attributes, but it is no longer worse than the reference. To assess where the attributes are relative to the reference or anchor, we need the third choice set. In this case, it is coded with -1s, indicating the attributes present, and a reference row. The reference row is chosen. The interpretation is that the reference is chosen only when the other attributes (that are all more important than the threshold from the dual response question) are negated or not present.

Had the respondent selected "none of these are important" in the dual response question the coding would be as in Figure 2B, but the -1s in the third choice set would be changed to 1s with the reference row still chosen. The interpretation would be that the reference row is chosen when the other attributes are present, or that all of the attributes are less important than the threshold.

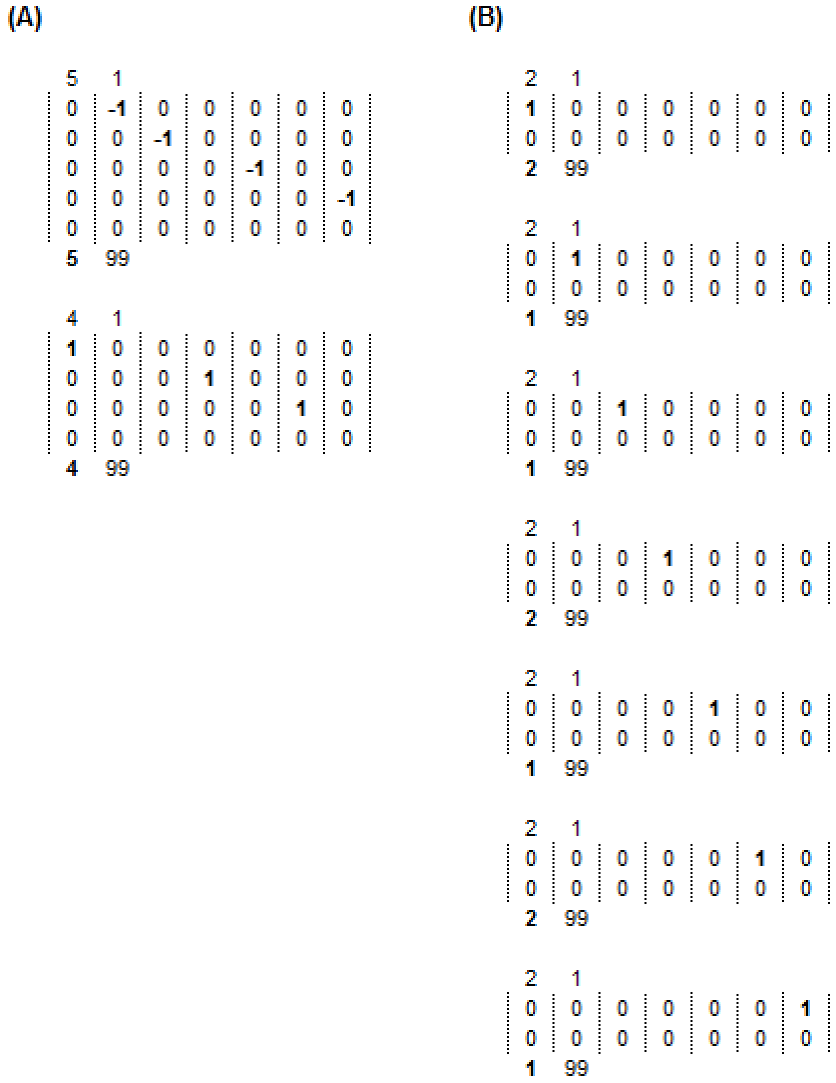
MaxDiff with Direct Binary Anchoring was coded as described in Lattery (2010), and using a method similar to one introduced at a recent SKIM/Sawtooth Software Conference by Bryan Orme (Orme, 2009a). Imagine the same situation as above where seven attributes are considered, but where attributes 2, 3, 5 and 7 are selected as "very important" in the direct binary question. The tradeoff screens are coded as in any ordinary (un-anchored) set of MaxDiff tasks (i.e., as shown in the first two choice sets in Figure 2B). This means that there is no reference or anchor row on any of the choice sets.

Following the Lattery method, hereafter referred to as the “current” method, two additional choice sets are included as shown in Figure 3B. The first of these codes the attributes selected (#2, 3, 5 and 7) in the direct binary question as -1s and includes a reference row. The second codes the unselected attributes (#1, 4, and 6) from the same question as 1s and also includes a reference row. In both sets, the reference row is the one selected. The interpretation from the first of these choice sets is that the reference is chosen when the attributes above it are negated (not present), and from the second set that the reference is chosen when the attributes above it are present.

Another way to code these same data, the “pairwise” method following Orme (2009a), is shown in Figure 3B. In this method, the MaxDiff tasks are coded as described above (i.e., as shown in the first two choice sets in Figure 2B). The direct binary responses are coded using one choice set for each attribute. Each of these choice sets contains a row indicating the attribute and one indicating the reference. If the respondent selects a given attribute in the direct binary question, the attribute row in the appropriate choice set is chosen; if the respondent does not select a given attribute in the same question, the reference row is chosen instead.

Direct Binary Only was coded using the current and the pairwise methods described above. Because there were no tradeoff exercises that preceded this condition only the choice sets associated with the direct binary questions were included (see Figure 3).

Paired Comparisons with Dual Response Anchoring was coded analogous to MaxDiff with Dual Response Anchoring. Imagine again the situation where seven attributes are considered and the respondent evaluates two of these (attributes 3 and 5) on a given choice screen. The respondent chooses attribute 3 as “best” and selects “one is important, the other is not” in the dual response question. This set of choices is coded as in Figure 4A. The first choice set indicates that attribute 3 is chosen over both attribute 5 and the reference or anchor row. The second choice set indicates that the reference is chosen when attribute 5 is present. Unlike MaxDiff with Dual Response Anchoring, this method results in a complete coding environment. We know from the responses and the coding where both attributes are relative to one another and relative to the reference.



(A)

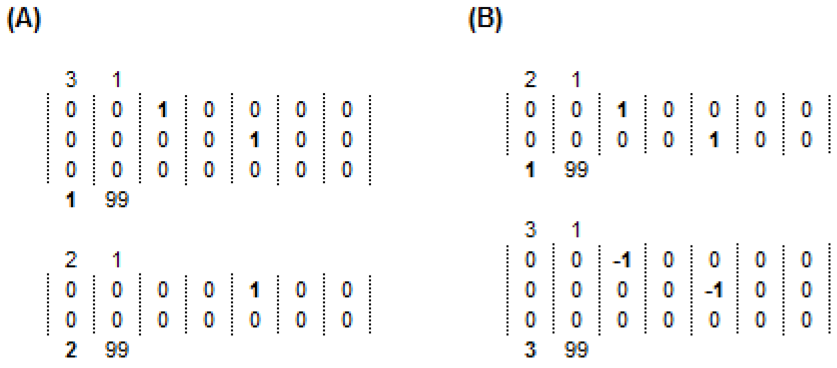
```
5   1
0  -1   0   0   0   0   0
0   0  -1   0   0   0   0
0   0   0   0  -1   0   0
0   0   0   0   0   0  -1
0   0   0   0   0   0   0
5  99

4   1
1   0   0   0   0   0   0
0   0   0   1   0   0   0
0   0   0   0   0   1   0
0   0   0   0   0   0   0
4  99
```

(B)

```
2   1
1   0   0   0   0   0   0
0   0   0   0   0   0   0
2  99

2   1
0   1   0   0   0   0   0
0   0   0   0   0   0   0
1  99

2   1
0   0   1   0   0   0   0
0   0   0   0   0   0   0
1  99

2   1
0   0   0   1   0   0   0
0   0   0   0   0   0   0
2  99

2   1
0   0   0   0   1   0   0
0   0   0   0   0   0   0
1  99

2   1
0   0   0   0   0   1   0
0   0   0   0   0   0   0
2  99

2   1
0   0   0   0   0   0   1
0   0   0   0   0   0   0
1  99
```

Figure 3. *.cho file coding for MaxDiff with direct binary anchoring, showing only the portions required for estimating the threshold relative to the attributes; attributes 2, 3, 5 and 7 selected in the direct binary grid question; **Panel A**: current method; **Panel B**: pairwise method.

Consider the same situation, but where the respondent selects “both are important” in the dual response question. This pattern is coded as in Figure 4B. The first choice set indicates only that attribute 3 is chosen over attribute 5 and omits the reference row. The second choice set is coded with -1s and indicates that the reference row is chosen when the two attributes are negated or not present. Had the respondent selected “neither of these are important” in the dual response question the second choice set would be coded with 1s instead of -1s and the reference row would again be chosen. The interpretation would be that the reference is chosen when the two attributes are present.



(A)

```
3  1
0  0  1  0  0  0  0
0  0  0  0  1  0  0
0  0  0  0  0  0  0
1  99

2  1
0  0  0  0  1  0  0
0  0  0  0  0  0  0
2  99
```

(B)

```
2  1
0  0  1  0  0  0  0
0  0  0  0  1  0  0
1  99

3  1
0  0  -1  0  0  0  0
0  0  0  0  -1  0  0
0  0  0  0  0  0  0
3  99
```

Figure 4. *.cho file coding for paired comparisons with dual response anchoring; 7 attributes in all; attributes 3 and 5 shown; attribute 3 chosen as “best;” **Panel A**: “one is important, the other is not;” **Panel B**: “both are important.”

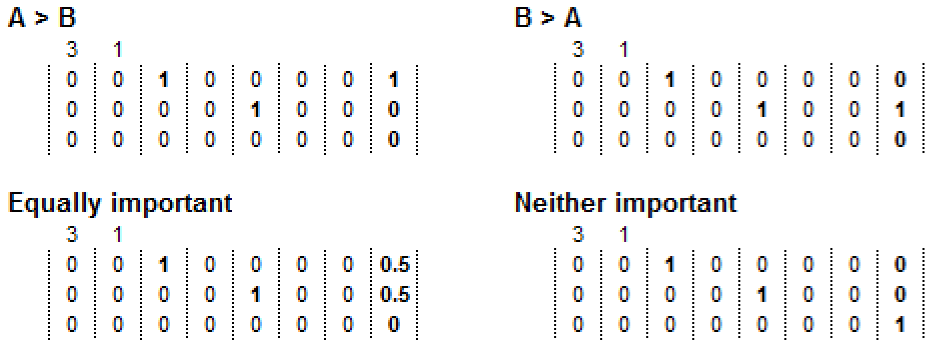
Paired Comparisons with Direct Binary Anchoring was coded exactly as in MaxDiff with Direct Binary Anchoring (see Figure 3) for the direct binary question using both the current and pairwise methods. The only differences were that the tradeoff choice sets that preceded those for the direct binary question consisted of only two attributes each; and there was only one set for each tradeoff task (“best” only), as opposed to two sets in MaxDiff (“best” and “worst”). An example of the paired comparisons tradeoff choice sets is shown in the first choice set in Figure 4B.

Two-by-Four shows two attributes per task and asks for a single response, given four options:

- Attribute A is more important than attribute B
- Attribute B is more important than attribute A
- Both are equally important
- Neither is important

The method is coded as an allocation task (*.chs file) in one of four ways depending on the response (Figure 5). In all cases, a single choice set is produced for each task that contains rows associated with both attributes and a reference row. When one of the first two response categories is selected, a point is allocated to the row that was selected as more important; when “both are equally important,” the point is split between the two attribute rows; and when “neither is important,” the point is allocated to the reference row.

The first two examples indicate that the item selected as more important is not only more important than the other but that it is more important than the reference. It does not however indicate that the attribute not selected is less important than the threshold (nor does the response imply that is the case). Since we do not have information about where the not selected attribute is relative to the reference, either from the response or the coding, this method is incomplete. When “both are equally important” is selected, we have complete coding, but perhaps an inference that is not correct. The coding suggests that the response indicates that both are more important than a fixed reference point; but since we have not asked that question directly, it might not be the case. Finally, when “neither is important” is selected, we lack complete coding again. We know where both attributes are relative to the reference, but we do not know where they are relative to one another. It seems that when presented with two attributes, a single response is not sufficient to fully account for all the relationships between the attributes and a threshold.



A > B

3	1						
0	0	1	0	0	0	0	1
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0

B > A

3	1						
0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	1
0	0	0	0	0	0	0	0

Equally important

3	1						
0	0	1	0	0	0	0	0.5
0	0	0	0	1	0	0	0.5
0	0	0	0	0	0	0	0

Neither important

3	1						
0	0	1	0	0	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1

Figure 5. *.chs coding for Two-by-Four; 7 attributes in all; attributes 3 and 5 shown.

Results

Correlations of HB Utilities Between Methods

Correlations among the HB-estimated logit utilities across the nine different methods are shown in Figure 6. The current and pairwise coding methods employed in the three conditions that used a direct binary anchor correlated perfectly, or nearly so, within condition (*r*=1, 1, and .97 for MaxDiff, paired comparisons and direct binary only, respectively). The pairwise coding method stretched the utilities relative to the current one in all three conditions. This was an expected result (Kevin Lattery, personal communication), albeit a small one.

The utilities from the two dual response conditions (MaxDiff and paired comparisons) also correlated very highly with the direct binary conditions within data collection type (*r* = .92). The differences in making these comparisons came from where the threshold was estimated, which can be seen in the scatterplots as standing out from the line of best fit in the two dual response columns. In both cases, the threshold was estimated lower in the dual response methods than in the direct binary ones. This finding was also in agreement with Lattery (2010). The effect appeared to be larger in the MaxDiff methods than in paired comparisons, but again the difference is small.

MaxDiff and paired comparisons results also correlated very highly with one another, but especially within coding method (*r*=.98). Correlations across data collection and anchoring methods fared only slightly worse (*r*=.90-.95). The poorest correlations among the conditions were between direct binary only and Two-by-Four, and between direct binary only and the two dual response methods.

Position of Threshold Varies by Method

Regardless of experimental condition, the attributes by themselves correlated highly with one another. It was the estimation of the threshold that varied. Figure 7 shows the rank order of the

HB-estimated logit utilities for all 17 attributes (light grey lines) and the threshold (bold line) across experimental conditions.

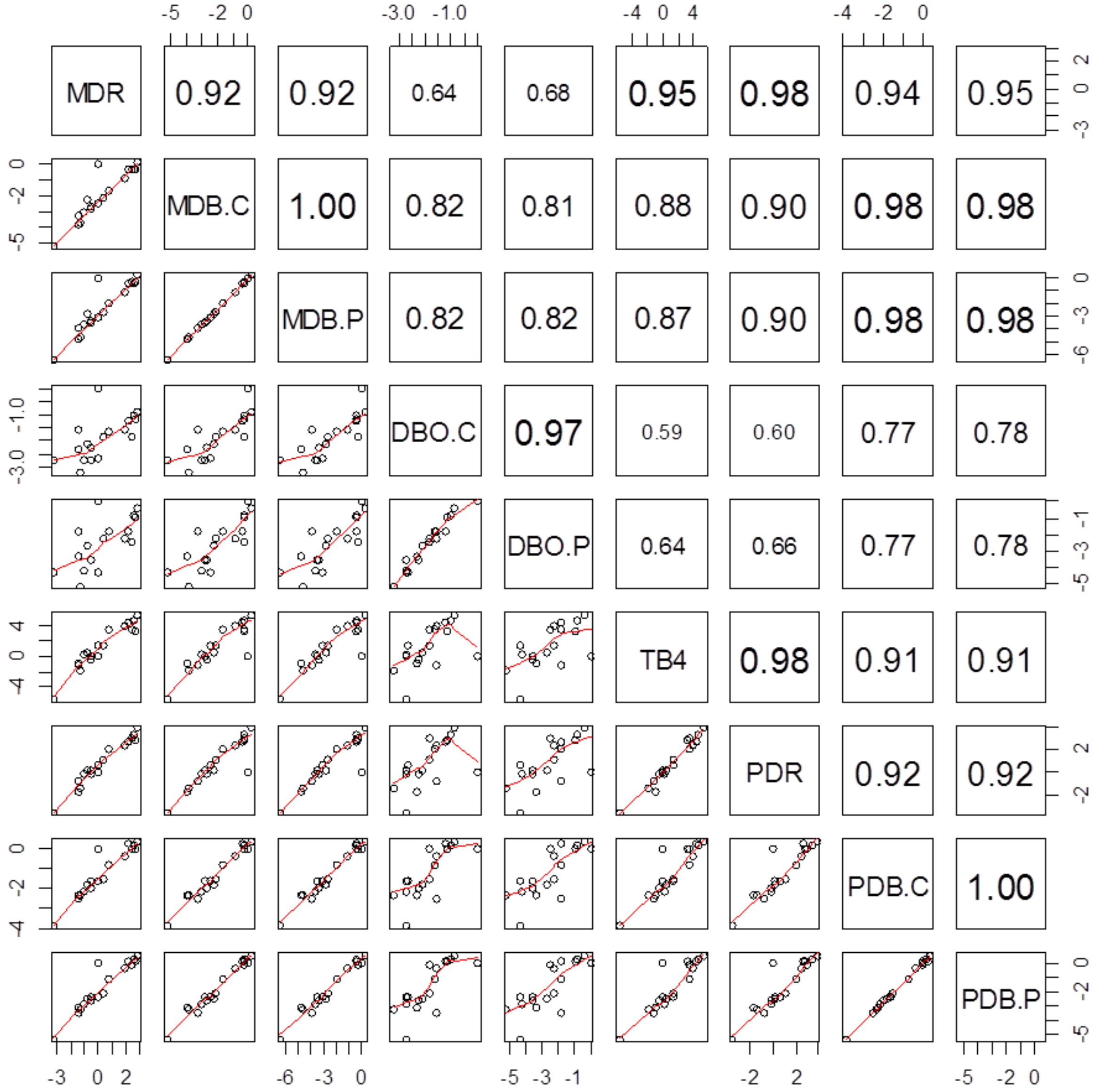


Figure 6. Correlations among utilities across methods. **MDR** = MaxDiff, Dual Response; **MDB.C** = MaxDiff, Direct Binary, Current Coding; **MDB.P** = MaxDiff, Direct Binary, Pairwise Coding; **DBO.C** = Direct Binary Only, Current Coding; **DBO.P** = Direct Binary Only, Pairwise Coding; **TB4** = Two-by-Four; **PDR** = Paired Comparisons, Dual Response; **PDB.C** = Paired Comparisons, Direct Binary, Current Coding; **PDB.P** = Paired Comparisons, Direct Binary, Pairwise Coding.

The relative rank order of the top-most and bottom-most attributes was particularly stable across conditions. Even towards the middle of the rank order, there were attributes that varied only a small amount. Two-by-Four may be a notable exception in its ranking of several attributes relative to other conditions. However, the threshold moved a great deal across experimental condition. In the direct binary only conditions, no attributes exceeded the threshold, while in

MaxDiff with dual response anchoring, all but three did. The position of the threshold was stable among the current and pairwise coding methods involving direct binary anchoring, moving only one or two places within either MaxDiff or paired comparisons (it was higher in the pairwise coding method both times); and the threshold was estimated in about the same place between MaxDiff and paired comparisons with direct binary anchoring.

The threshold was quite a bit higher in the rank order in paired comparisons with dual response anchoring than it was in MaxDiff with dual response anchoring. This may have been due to the different psychological experience when being asked the dual response question when two items are present compared to when there are four items. It may be easier for the respondent to say "both items are important" than it is to say "all 4 items are important." Likewise for "neither of these items are important" compared to "none of these 4 items are important." Following this logic, we would expect a larger proportion of respondents to have answered "some are important, some are not" in the MaxDiff condition than those who answered "one is important, the other is not" in the paired comparisons condition, and this is, in fact, what occurred (Table 1). Curiously, the odds ratio of the first two responses was about the same between MaxDiff and paired comparisons (1.74 vs. 1.65), but the third response was selected nearly twice as often in MaxDiff than it was in paired comparisons.

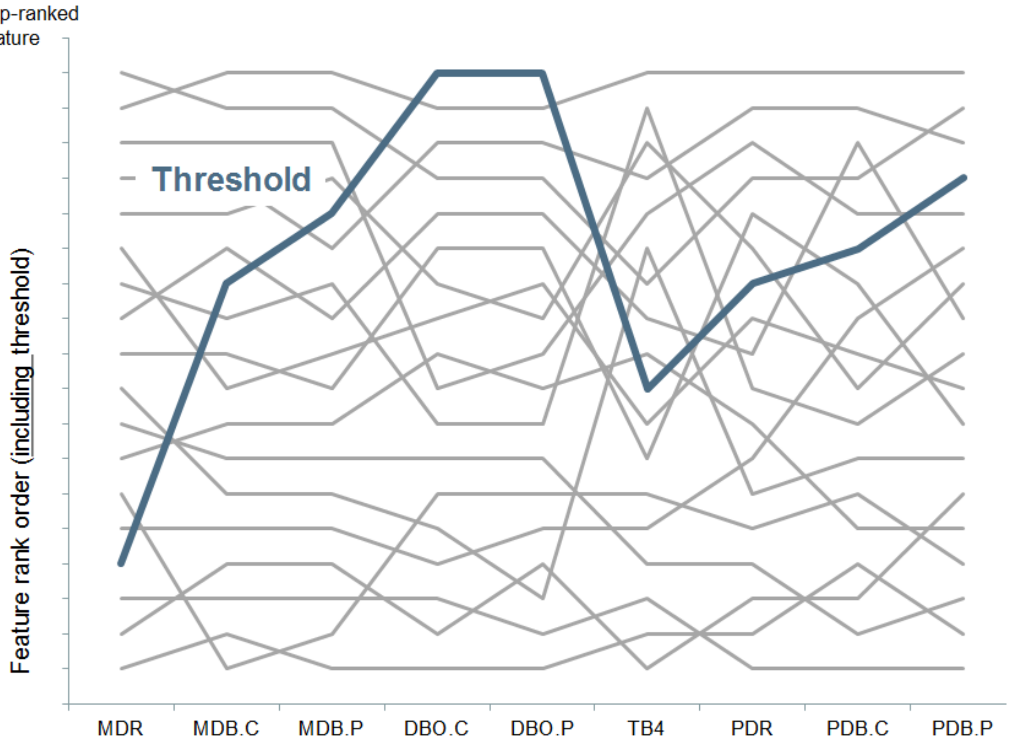
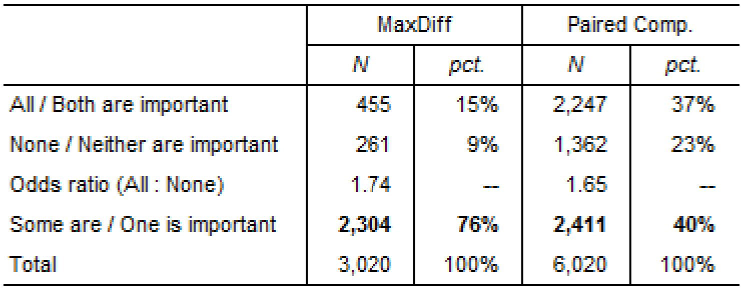


Figure 7. Relative ranks of 17 **attributes** (light grey lines) and **threshold** (bold line) across methods. Ranks are based on HB analyses. **MDR** = MaxDiff, Dual Response; **MDB.C** = MaxDiff, Direct Binary, Current Coding; **MDB.P** = MaxDiff, Direct Binary, Pairwise Coding; **DBO.C** = Direct Binary Only, Current Coding; **DBO.P** = Direct Binary Only, Pairwise Coding; **TB4** = Two-by-Four; **PDR** = Paired Comparisons, Dual Response; **PDB.C** = Paired

Comparisons, Direct Binary, Current Coding; **PDB.P** = Paired Comparisons, Direct Binary, Pairwise Coding.

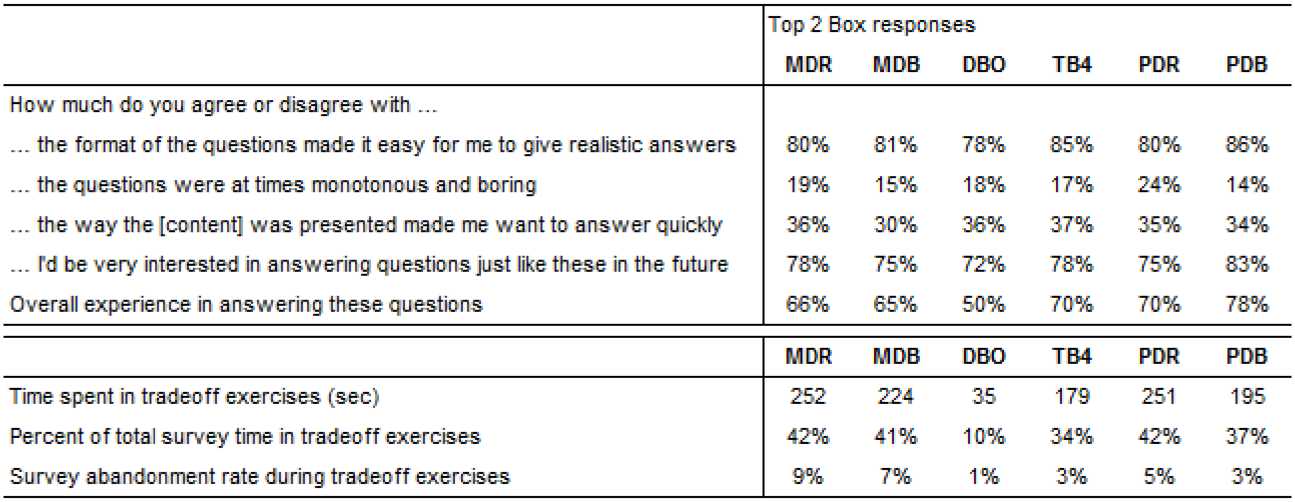


	MaxDiff		Paired Comp.	
	N	*pct.*	*N*	*pct.*
All / Both are important	455	15%	2,247	37%
None / Neither are important	261	9%	1,362	23%
Odds ratio (All : None)	1.74	--	1.65	--
Some are / One is important	**2,304**	**76%**	**2,411**	**40%**
Total	3,020	100%	6,020	100%

Table 1. Frequency table of dual response questions associated with MaxDiff (4 items per task) and paired comparisons (2 items per task).

Self-Evaluation and Abandon Rates

Paired comparisons with direct binary anchoring was the condition viewed most favorably by respondents (Table 2). That condition, along with Two-by-Four had two of the lowest abandon rates during the tradeoff tasks. Abandon rates were 2% lower in both MaxDiff and paired comparisons when direct binary anchoring was used, compared to when dual response anchoring was used in the same conditions. This may have resulted from the shorter time required to complete the former anchoring method.



	Top 2 Box responses					
	MDR	**MDB**	**DBO**	**TB4**	**PDR**	**PDB**
How much do you agree or disagree with ...						
... the format of the questions made it easy for me to give realistic answers	80%	81%	78%	85%	80%	86%
... the questions were at times monotonous and boring	19%	15%	18%	17%	24%	14%
... the way the [content] was presented made me want to answer quickly	36%	30%	36%	37%	35%	34%
... I'd be very interested in answering questions just like these in the future	78%	75%	72%	78%	75%	83%
Overall experience in answering these questions	66%	65%	50%	70%	70%	78%
	MDR	**MDB**	**DBO**	**TB4**	**PDR**	**PDB**
Time spent in tradeoff exercises (sec)	252	224	35	179	251	195
Percent of total survey time in tradeoff exercises	42%	41%	10%	34%	42%	37%
Survey abandonment rate during tradeoff exercises	9%	7%	1%	3%	5%	3%

Table 2. Summary of responses from self-evaluation questions, amount of time spent in the tradeoff tasks and abandonment rates across methods. **MDR** = MaxDiff, Dual Response; **MDB** = MaxDiff, Direct Binary; **DBO** = Direct Binary Only; **TB4** = Two-by-Four; **PDR** = Paired Comparisons, Dual Response; **PDB** = Paired Comparisons, Direct Binary.

EXPERIMENT 2: SIMULATIONS

Methods

Simulations were used to determine which of the nine methods used in Experiment 1 best recover actual utilities. The following factors were tested, resulting in 270 simulated data sets:

- 9 methods (data collection methods and coding variations
- 3 sample sizes (75, 150, 300)
- 2 levels of attributes (10, 20)
- 5 replicates

Actual utilities ranged from -4.5 to 0 in equal increments, depending on the number of attributes. Right-skewed Gumbel error (loc=0, scale=1) was added to actual utilities for “best” choices, and left-skewed Gumbel error was added for “worst” choices. Methods and actual utilities generally followed those reported by Orme (2007). The threshold value was set at 0 for all simulations. Placing the threshold at the upper end of the distribution likely advantages dual response methods over the direct binary ones. Had we placed the threshold in the center of the distribution, the reverse would have been true. Additional research is in progress to determine whether the placement of the threshold in simulations affects utility recovery, and the degree to which varying the error assumptions for the responses used to estimate the threshold affects utility recovery.

Choice data were generated from the individual-level utility matrices and design files for each experimental condition. These were recoded into appropriate *.cho or *.chs files. Data were analyzed using aggregate logit models.

Results

Mean absolute error (MAE) was smallest in the paired comparisons with dual response anchoring method (Figure 8), followed closely by MaxDiff and paired comparisons, both with direct binary anchoring and pairwise coding. The pairwise coding method cut MAE in half on average relative to the current method in all three conditions where both coding methods were used.

The worst performers in terms of MAE were direct binary only and Two-by-Four. As a result of this poor performance, and some incomplete coding issues in the Two-by-Four method discussed earlier, we do not recommend the use of either of these methods over MaxDiff or paired comparisons to discover the relative value among items.

Figure 9 further demonstrates the smaller MAE associated with the pairwise coding method of direct binary responses. When the current coding method was employed, there was an upward bias on the estimated utilities relative to the actual ones. This bias was especially prevalent towards the upper end of the scale. The same bias was apparent when the pairwise coding method was used, but was less pronounced. These biases may indicate that the responses to the direct binary question follow a different error distribution, yet to be tested. Nevertheless, these results appear to favor the pairwise coding method of direct binary responses.

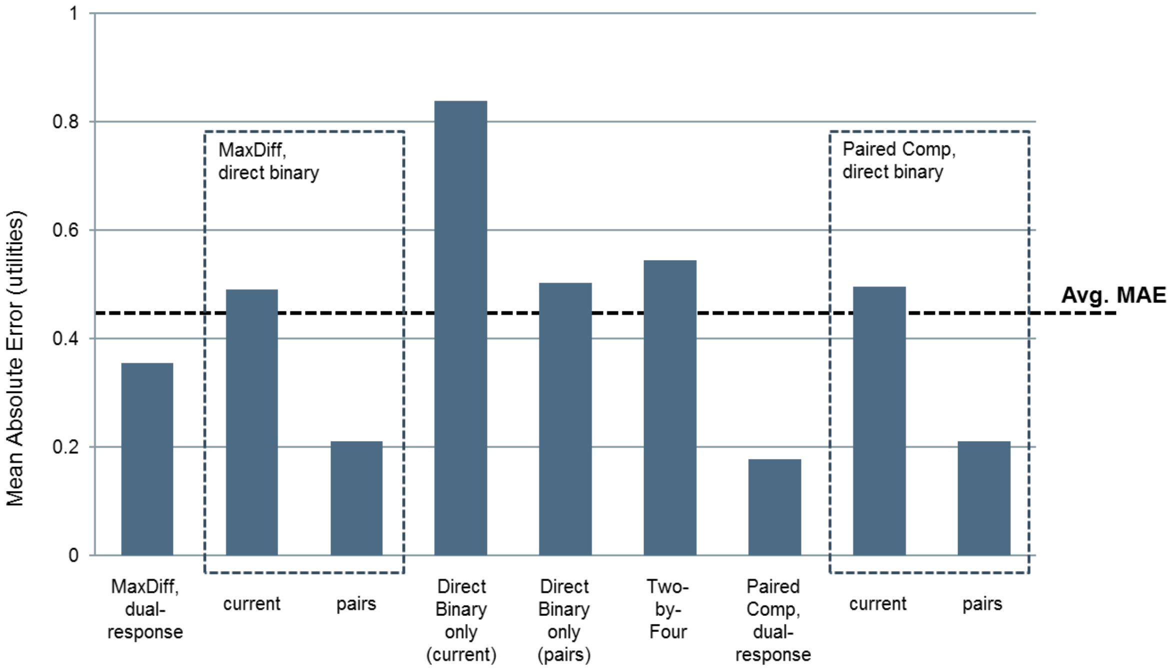


Figure 8. Mean absolute error (estimated utilities vs. actual) across methods.

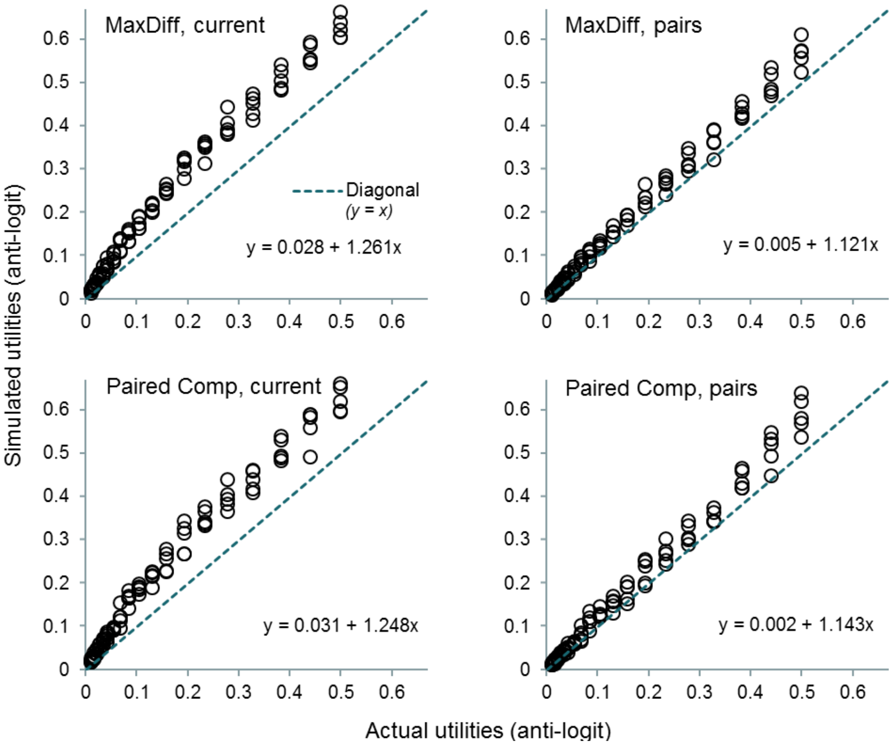


Figure 9. Recovery of utilities from tradeoffs with direct binary anchoring, comparing current and pairwise coding methods.

Meta-Analysis: Scale-Use Bias?

Results from a large multi-country experiment using MaxDiff with direct binary anchoring suggest that the method may re-introduce scale-use biases. Figure 10 is one example from this meta-analysis that involved 21 different sets of attributes (ranging in number from 8 to 20) measured across six different countries with more than 22,000 respondents. Thresholds were typically lower among the rank orders in BRIC countries (Brazil, Russia, India and China) than they were in the US and Germany. This resulted from a greater tendency of respondents in BRIC countries to select items in the direct binary questions than US or German respondents.

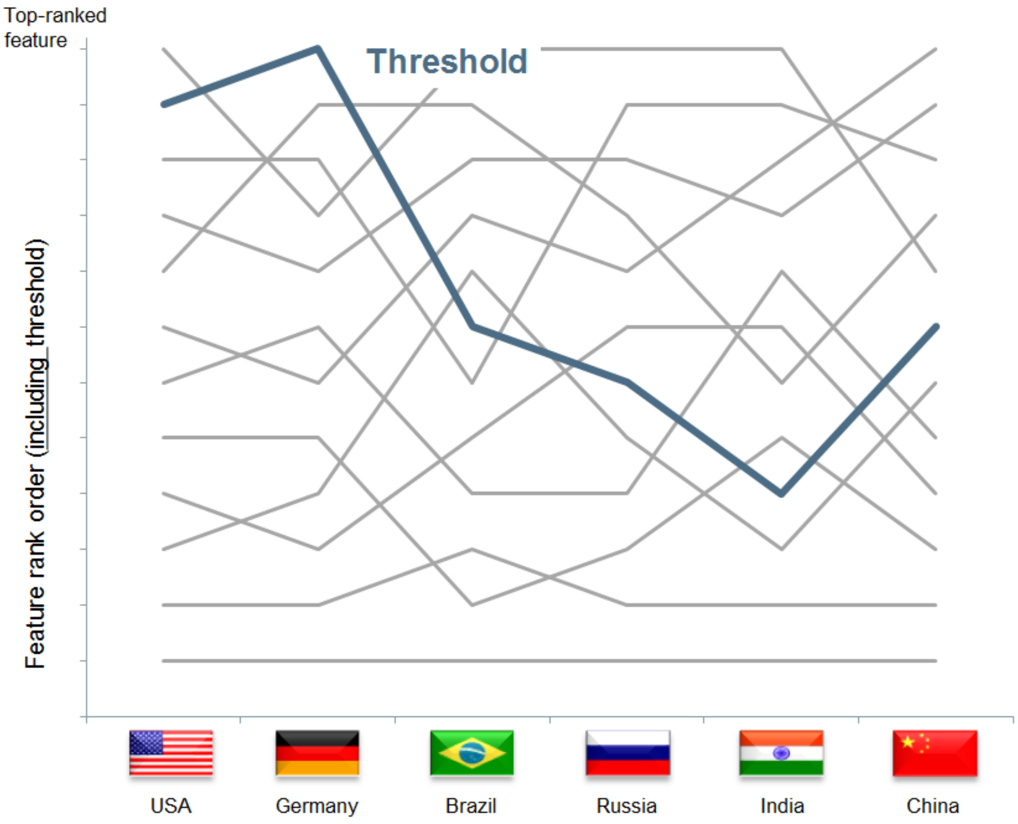


Figure 10. Relative ranks of the same 11 **attributes** (light grey lines) and **threshold** (bold line) estimated separately in six different countries. Ranks are based on HB analyses.

Discussion and Conclusions

Paired Comparisons or MaxDiff?

Respondents in Experiment 1 completed 20 paired comparisons tasks in slightly less time than they completed 10 MaxDiff tasks (avg. 223 sec compared to 238 sec). Favorability ratings for paired comparisons also exceeded those of MaxDiff. Recovery of actual utilities was about the same between the methods. In addition to these observations, we can also note that paired comparisons with dual response anchoring has an associated complete coding method, while

MaxDiff with dual response anchoring does not. These findings suggest that we should not be afraid to use paired comparisons, especially when the attributes being tested span a sentence or two, even when the number of tasks required might run to 20 or more.

These findings however also appear to conflict with those reported by Cohen and Orme (2004). They found MaxDiff results to have better predictive validity of holdout questions than paired comparisons. (We did not seek to predict holdout tasks from our results. Due to the large number of variations tested and different numbers of tasks between methods, it would have been difficult to assess predictive validity from holdouts in a consistent way across methods. Therefore, it is not possible to determine whether there is a real conflict here, or whether we are simply measuring two different things and arriving at different and equally-valid results as a consequence.) MaxDiff, by placing more than two attributes in each choice task, increases the number of *implied* paired comparisons. In any group of four items, there are six theoretical paired comparisons, five of which are known from choosing a "best" and a "worst" item. MaxDiff analysis of 10 choice tasks results in 50 implied paired comparisons, which is 2.5 times as much information as is in 20 paired comparison choice tasks. This may be the reason that Cohen and Orme (2004) found MaxDiff to have greater predictive validity.

In the final analysis, both MaxDiff and paired comparisons should be given their proper due. Paired comparisons may be better for a smaller number of wordy attributes, even as the number of tasks grows to be uncomfortably large (as many as 20). MaxDiff is a good way to test a larger number of features without overwhelming respondents. We do not yet know whether the results reported here vis-à-vis paired comparisons will still apply with 25, 30 or 35 items. MaxDiff may prove superior as the total number of items grows.

Anchoring and Coding Methods

Direct binary methods involved less respondent burden, were completed faster and viewed more favorably than dual response methods. Further, when using MaxDiff exercises, both coding methods described here for direct binary responses are mostly complete (we still do not know the relationship between the two not selected items); dual response coding is incomplete.

These are all arguments in favor of using the direct binary anchoring method, especially in the context of MaxDiff exercises. However, the direct binary question may be more sensitive to context effects, especially as the number of attributes increases (an attribute list containing 8 items is easier to evaluate in a single grid question than is a list of 20). Bryan Orme tells us that "the old MaxDiff and paired comparisons were lacking an absolute anchor … but they were free from scale use bias! When we introduce the idea of anchored scaling, all these approaches let that 800lb gorilla back in the room." (personal communication, ellipsis in original). Meta-analysis of cross-cultural data presented here suggests that this concern over scale-use bias is valid. Whether one or another of these anchoring methods is more resistant to these biases remains to be investigated. Nevertheless, we have shown that the biases do exist, at least for the direct binary method used with MaxDiff.

Finally, when the direct binary anchoring method is used, the results presented here suggest that the pairwise coding method is superior to the current one, at least for recovering actual utilities. This conclusion assumes that we have sampled the correct error distribution in our simulations, which is a question we continue to investigate. This pairwise coding method increases the number of choice tasks in the *.cho file, so in a way, it is not surprising to find better recovery of utilities. However, there is not really more information in a *.cho file resulting from the pairwise method than there is in one from the current method. The pairwise coding method also seems to mimic better what respondents are doing (especially if we present respondents with a two-point scale for each attribute). This, in the end, provides some justification for adopting the pairwise coding method.

References

Bacon, L., Lenk, P., Seryakova, K. and Veccia, E. (2007). Making MaxDiff More Informative: Statistical Data Fusion by way of Latent Variable Modeling. *2007 Sawtooth Software Conference Proceedings*, Santa Rosa, CA.

Chrzan, K. (2004). The Options Pricing Model: A Pricing Application of Best-Worst Measurement. *2004 Sawtooth Software Conference Proceedings*, San Diego, CA.

Cohen, S. (2003). Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation. *2003 Sawtooth Software Conference Proceedings*, San Antonio, TX.

Cohen, S. and Orme, B. (2004). What's Your Preference? *Marketing Research*, 16, pp.32-37.

Johnson, P. and Fuller, B. (2012). Optimizing Pricing of Mobile Apps with Multiple Thresholds in Anchored MaxDiff. *2012 Sawtooth Software Conference Proceedings*, Orlando, FL.

Lattery, K. (2010). Anchoring Maximum Difference Scaling Against a Threshold – Dual Response and Direct Binary Responses. *2010 Sawtooth Software Conference Proceedings*, Newport Beach, CA.

Louviere, J. J. (1991). Best-Worst Scaling: A Model for the Largest Difference Judgments. Working Paper, University of Alberta.

Magidson, J., Thomas, D. and Vermunt, J.K. (2009). A New Model for the Fusion of MaxDiff Scaling and Ratings Data. *2009 Sawtooth Software Conference*, Delray Beach, FL.

Orme, B. (2007). MaxDiff/Web v6.0 Technical Paper. Available at http://www.sawtoothsoftware.com/education/techpap.shtml

Orme, B. (2009a). Using Calibration Questions to Obtain Absolute Scaling in MaxDiff. *SKIM/Sawtooth Software European Conference. Prague*, Czech Republic.

Orme, B. (2009b). Anchored Scaling in MaxDiff Using Dual-Response. Available at http://www.sawtoothsoftware.com/education/techpap.shtml

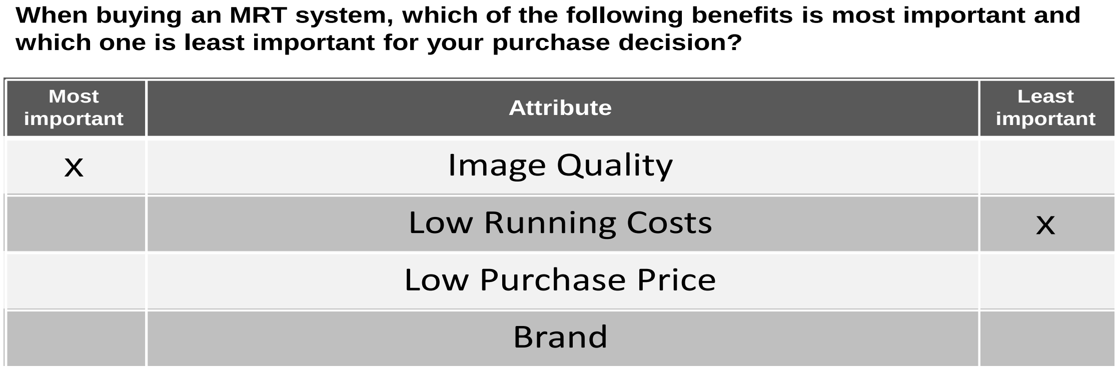
USING MAXDIFF FOR EVALUATING VERY LARGE SETS OF ITEMS

RALPH WIRTH
ANETTE WOLFRATH
GFK MARKETING SCIENCES

INTRODUCTION

Deficiencies of rating scales for measuring attribute importances and preferences are well known. In particular, rating scales suffer from response style bias (e.g. Baumgartner, Hofstede 2001; Rossi, Gilula, Allenby 2001), lack of discrimination between items (e.g. Chrzan, Golovashina 2006; Wirth 2008), low test-retest reliability (Wirth 2008) and weak predictive power (e.g. Bass, Wilkie 1973; Beckwith, Lehmann 1973, Wilkie, Pessemier 1973). These deficiencies lead to numerous serious problems in such popular analyses as multinational segmentations and the identification of purchase key drivers.

In recent years, a scaling method called "Maximum Difference Scaling" (MaxDiff Scaling), often also called "Best-Worst Scaling" (Finn/ Louviere, 1992) has attracted considerable attention among researchers and practitioners (e.g. Marley/ Louviere, 2005; Cohen, 2003; Cohen/ Orme, 2004; Chrzan, 2006). It has been shown that Best-Worst Scaling does not suffer from typical weaknesses of traditional rating scales, as it leads to greater discrimination among items (Cohen, 2003; Cohen/Orme, 2004; Chrzan, 2006, Wirth 2008), greater predictive power (Chrzan, Golovashina 2006), greater test-retest reliability and fewer inconsistencies in observed response behavior (Wirth 2008).



When buying an MRT system, which of the following benefits is most important and which one is least important for your purchase decision?

Most important	Attribute	Least important
x	Image Quality	
	Low Running Costs	x
	Low Purchase Price	
	Brand	

Figure 1
Example MaxDiff Task

However, there is one major limitation that makes the application of the Best-Worst scaling approach difficult in some situations: If useful individual-level scores are needed, the number of required MaxDiff tasks (similar to the one shown in Figure 1) that each respondent has to complete increases with the total number of items to be studied. According to a common rule of thumb that each item should be shown at least three times to each respondent (for stable individual-level scores), a study involving 30 items would already result in at least 22 four-item MaxDiff tasks or at least 18 five-item MaxDiff task. While this may still seem feasible, studies dealing with 60 items require at least 45 four-item MaxDiff tasks or at least 36 tasks showing five items each. Most researchers would agree that these settings pose too high a burden on the respondents. Hendrix' and Drucker's (2008) observation, that

> *"[a] review of the literature on empirical use of MaxDiff in marketing research consistently presents analysis of 30 or fewer items at a time"*

confirms that the fear of overburdening respondents clearly limits the scope of MaxDiff studies conducted by practitioners.

Although studies dealing with 100 items or more are certainly not the rule, they are no exception either. Therefore, certain workarounds, such as aggregate MaxDiff models (where stable individual-level models is not the aim) or more traditional approaches to measuring attribute importances, are often used in these cases. Obviously, this is not optimal. For that reason, researchers have tried to find ways to deal with larger item sets in MaxDiff studies, while still obtaining robust individual-level scores. Two particularly appealing approaches were introduced by Hendrix and Drucker (2008). Both methods are characterized by a combination of Q-Sort information and information from sparse MaxDiff experiments: Q-Sort information is either used for augmenting MaxDiff information during HB estimation ("Augmented MaxDiff"), or for individually tailoring the MaxDiff questionnaire on-the-fly ("Tailored MaxDiff"). The authors demonstrate in their paper that both approaches work well with 40 items.[3] Nevertheless, two concerns motivated us to examine another approach to MaxDiff for large item sets:

- Q-Sort tasks may become very cumbersome for very many items (such as e.g. 100),
- both Augmented and Tailored MaxDiff require solid programming skills, making the approaches inapplicable for many practitioners.

In the following, we describe the idea behind this alternative approach that we call "Express MaxDiff". The performance of this approach is tested using both a Monte Carlo simulation study and an empirical comparison study. The setup and the results of these studies will be explained in later sections.

"EXPRESS MAXDIFF": BASIC IDEA AND APPLICATION OF THE APPROACH

In order to reduce interview complexity or length, split questionnaire survey techniques are often used in market research studies (Rässler, Koller, Mäenpää, 2002). Split designs mean that respondents are confronted only with a (potentially small) subset of the questions during the interview. Across the entire sample, information is gathered for all questions, but with structural missing values. Commonly, imputation algorithms are used to fill in the missing values in order to receive a complete data matrix.

The idea behind split sample questionnaires is also the main characteristic of the Express-MaxDiff approach. In order not to overburden respondents in studies involving very many items, different *design blocks* are generated using an experimental design algorithm. Each

[3] In the meantime, more Augmented MaxDiff studies have been successfully conducted with larger numbers of items (personal correspondence with Drucker, 2012).

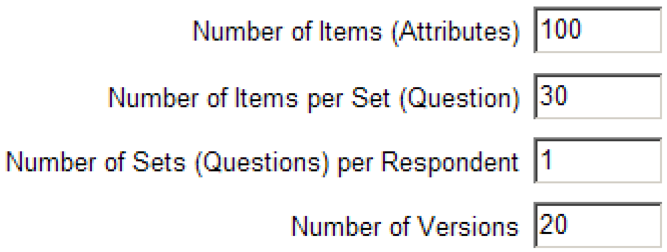
[4] We approached the estimation slightly differently, as we coded the design matrices in terms of utility differences (see explanations in the section

design block contains only a subset of the items to be studied, and only these subsets of items are used to create standard MaxDiff questionnaires and questionnaire versions, e.g. by using Sawtooth Software's MaxDiff designer. Of course, each respondent proceeds through one specific MaxDiff questionnaire. That means, that while over the whole sample all items are evaluated, each respondent actually only evaluates a subset of the items.

In order to get full parameter vectors for each individual anyway, Express-MaxDiff makes use of Hierarchical Bayes' (HB) "borrowing strength" characteristic. This widely used expression reflects the fact that HB approaches enable the estimation of individual parameters in spite of sparse information at the individual level by "borrowing" information from the rest of the sample. The estimation of individual parameters based on sparse Choice-Based Conjoint (CBC) data has mainly become possible due to of the advent of HB in market research at the end of the last century (see e.g. Allenby, Rossi 1999; Allenby, Rossi 2003; Allenby et al. 2004). In the case of Express-MaxDiff, individual information is relatively strong about items that have actually been evaluated, but there is *no* individual information about many other parameters. The information about these latter parameters is thus completely inferred from the information about the parameters of the rest of the sample. It is important to note that this does not imply a simple imputation of the sample average! Rather, information about covariance and uncertainty are taken into account during estimation. A simplified example: If it is known that, across the whole sample, respondents who like Item A also like Item C, this information is implicitly used when estimating individual parameters for Item C for those respondents who have only evaluated Item A.

For actually *designing MaxDiff experiments and estimating individual parameters using HB,* standard software can be used. The design process comprises two steps:

In the first step, the *design blocks* are generated using Sawtooth Software's MaxDiff Designer. For example, if there are 100 items in total, and the researcher wants to generate 20 design blocks containing 30 items each, MaxDiff Designer can be used to generate these 20 design blocks by entering the following figures while allowing individual designs lacking connectivity.



Number of Items (Attributes) 100

Number of Items per Set (Question) 30

Number of Sets (Questions) per Respondent 1

Number of Versions 20

Figure 2
Example Design Settings for Generating Design Blocks Using the MaxDiff Designer

Each version is then interpreted as a design block that contains only those items that are the respective "alternatives" in this task.

In the next step, a *standard MaxDiff experiment* is generated for each design block, once again using the Experimental Designer. That is, the researcher specifies the number of tasks and alternatives per tasks as well as the number of questionnaire versions per block and creates the experimental designs.

Before fielding the study using SSI Web, the item numbers in the questionnaire versions (which, in our example always range from 1 to 30 in all tasks) must be *replaced* by the actual item numbers (which, in our example, range from 1 to 100). This "translation" can be done using standard software, such as R.

When following this process, the analysis, i.e. the HB-estimation of individual parameters, can easily be done using CBC/HB, just like in normal MaxDiff studies.[4]

RESEARCH DESIGN

As mentioned before, the performance of Express-MaxDiff is evaluated using both a Monte-Carlo simulation study and an empirical comparison study.

Monte Carlo simulation studies are often used for assessing the general performance of new methods without having to rely on particular empirical data sets. The main idea behind simulation studies is the following: The researcher generates observations for a sample of "simulated" respondents based on "true" known parameters. This is done for many different experimental conditions, reflecting e.g. varying numbers of respondents, varying numbers of parameters, etc. For all experimental conditions, the generated observations are then used for estimating parameters as usual. By comparing true parameters and estimated parameters, the researcher can assess the performance of the whole approach.

In our concrete research project, the Monte Carlo simulation study mainly aims at

a. assessing whether Express-MaxDiff is generally able to recover true parameters in spite of the sparse individual information,
b. identifying success factors for Express MaxDiff studies.

Simulation studies are very valuable in terms of giving more "general" insights into weaknesses and strengths of an approach, as the results do not depend on few empirical data sets. Nevertheless, a good performance in simulation studies is necessary, but not sufficient for real-world superiority. For that reason, our research design also comprises an *empirical study*, which aims at comparing the performance of Express-MaxDiff with the performance of Augmented MaxDiff (Hendrix, Drucker 2008) and that of another possible approach to MaxDiff with large item sets.

Both studies will be introduced in the following section.

[4] We approached the estimation slightly differently, as we coded the design matrices in terms of utility differences (see explanations in the section about the simulation study), but both approaches work technically.

MONTE-CARLO SIMULATION STUDY

The process of the simulation study is illustrated in Figure 3.

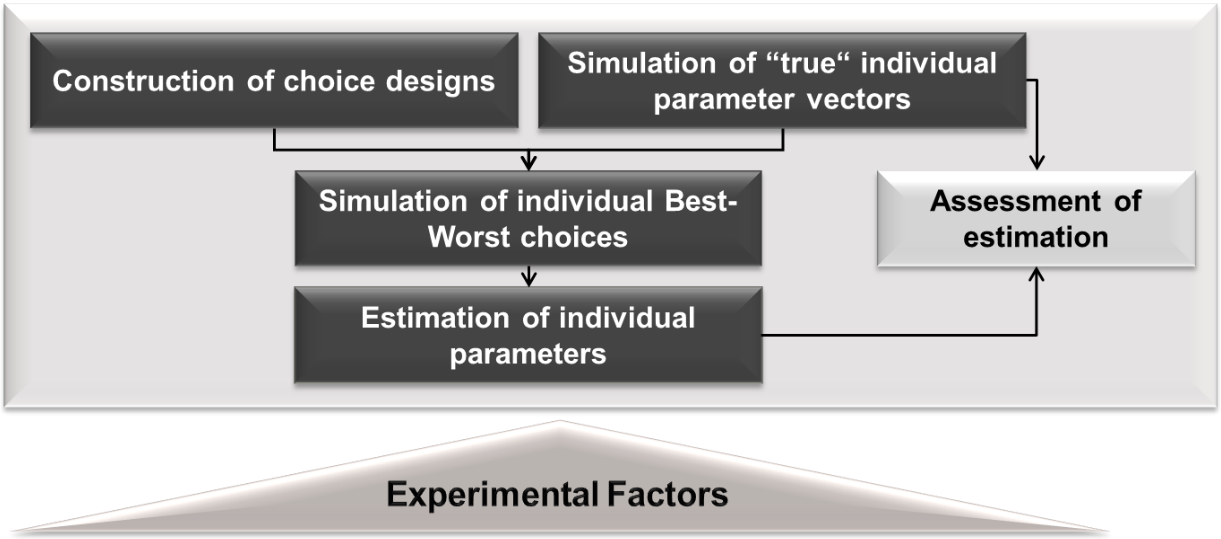
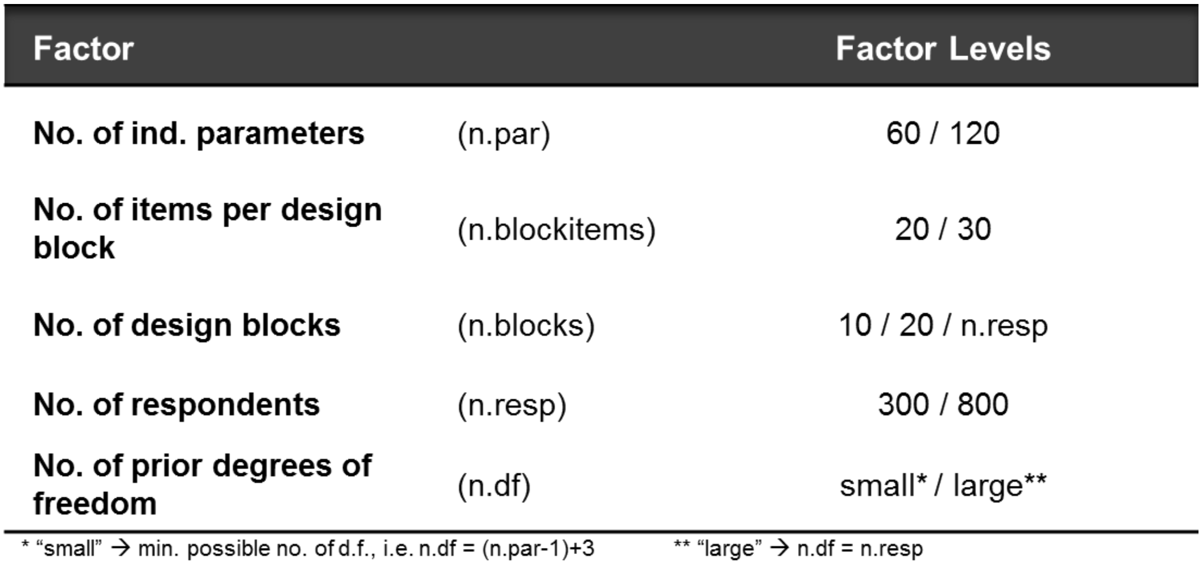

Construction of choice designs
Simulation of “true“ individual parameter vectors
Simulation of individual Best-Worst choices
Assessment of estimation
Estimation of individual parameters
Experimental Factors

Figure 3
Process of Simulation Study

The illustrated process follows the general explanations of Monte Carlo simulation studies in the preceding section: For a number of experimental conditions – defined by the experimental factors – choice data (i.e. MaxDiff-choices for a set of simulated respondents) are derived based on true individual preference parameters and choice designs, which are both generated by the researcher. These data are used for estimating preference parameters. By comparing true and estimated parameters, the performance of the approach can be assessed. Of course, the more similar true and estimated parameters are under a certain experimental condition, the better the approach works in this situation. Details about this process are explained in the following.

EXPERIMENTAL FACTORS AND DATA GENERATION

The experimental factors taken into account in the simulation study are illustrated in Table 1. This design leads to $2^4 \cdot 3 = 48$ different experimental situations that differ in terms of how challenging they are for the Express MaxDiff approach. For example, situations characterized by 120 individual parameters to be estimated but only 20 items per design block (i.e. only 20 items that are actually seen and evaluated by respondents) are more challenging than situations in which 60 total individual parameters have to be estimated based on design blocks containing 30 items each. The prior degrees of freedom are a parameter in HB analyses that influence the amount of information that is borrowed from the whole sample when estimating individual parameters. *Ceteris paribus*, the larger the number of prior degrees of freedom, the more important the sample information becomes.



Factor		Factor Levels
No. of ind. parameters	(n.par)	60 / 120
No. of items per design block	(n.blockitems)	20 / 30
No. of design blocks	(n.blocks)	10 / 20 / n.resp
No. of respondents	(n.resp)	300 / 800
No. of prior degrees of freedom	(n.df)	small* / large**

* “small” → min. possible no. of d.f., i.e. n.df = (n.par-1)+3 ** “large” → n.df = n.resp

Table 1
Experimental Factors and Factor Levels

For each of the 48 experimental situations, two data sets are generated by

1. creating the Express-MaxDiff design (i.e. design blocks and concrete MaxDiff tasks), then
2. generating individual parameter vectors that reflect the preferences of the artificial respondents, and finally
3. simulating the Best-Worst choices of the artificial respondents by “bringing together” the MaxDiff-tasks (step 1) and the individual preferences (step 2).

While the design generation has been described above, the generation of the individual parameters and the actual choices closely follows the procedure of simulating Best-Worst choices that is explained in Wirth (2010, 187-198) and Wirth (2011, 336-340).

Estimation and Performance Evaluation

Figure 4 illustrates how the performance of Express-MaxDiff is evaluated.

For each of the 96 simulated data sets, individual parameters are estimated using a standard HB-approach. We used Sawtooth Software’s CBC/HB software, but we modified the input design matrices in order to reflect the theoretically correct formulation of the MaxDiff likelihood function in terms of utility differences (see Marley, Louviere 2005; Wirth 2011, 324). However, researchers who are used to the more pragmatic approach of separating each MaxDiff task into two independent choice tasks (the “best”-choice task and the “worst”-choice task) can apply the same logic to Express MaxDiff without any problem.

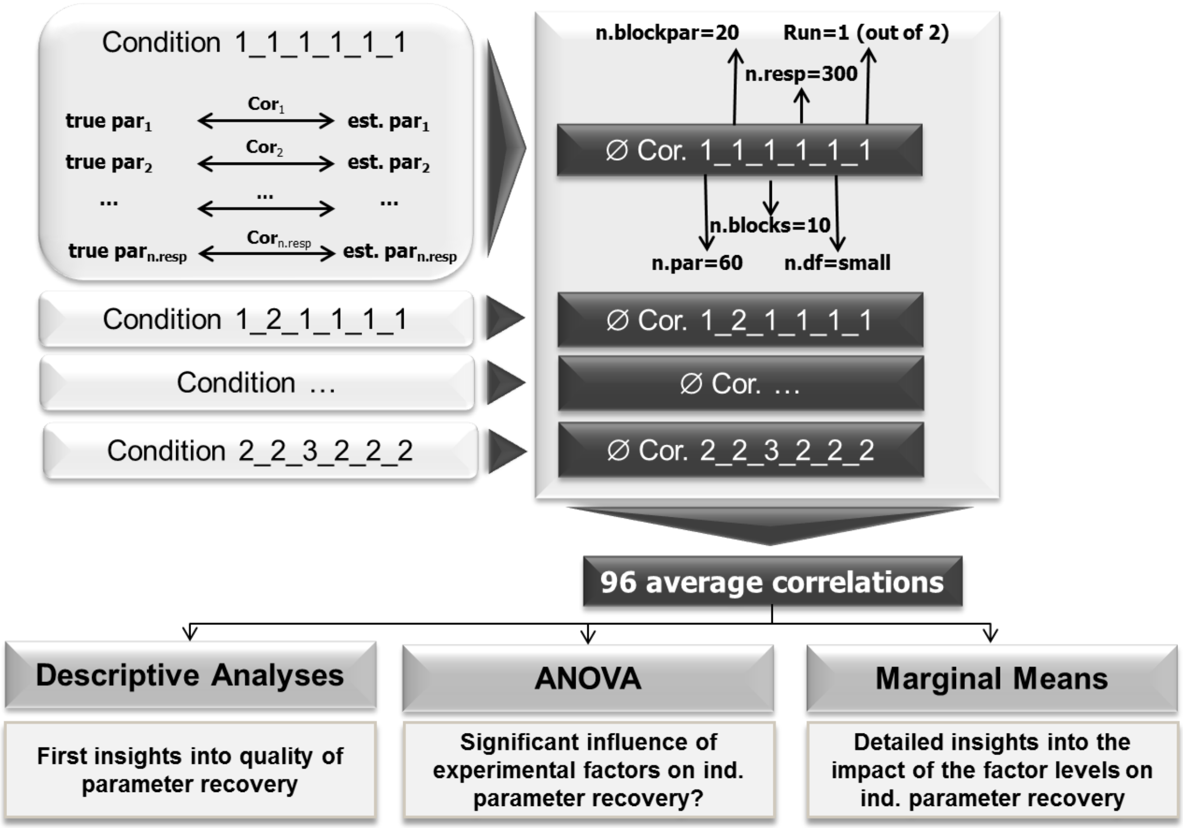


Figure 4
Performance Evaluation

The agreement between estimated and true individual parameters is measured using the Pearson correlation. For each data set, the n.resp individual correlations are averaged[5] in order to get a measure of how well true individual parameters are recovered under this specific experimental condition. These 96 average correlations are the basis for subsequent analyses: Descriptive analyses are used in order to get a first impression of the performance of Express MaxDiff; ANOVAs and t-tests give insight into the influence of the factors and factor levels on the goodness of the recovery of true individual parameters. The most important results of these analyses are explained in the following.

RESULTS

We start our analyses by looking at how well *aggregate* parameters are recovered when using the Express-MaxDiff approach. To do that, correlations between average true parameters and average estimated parameters are calculated for all 96 data sets. The results reveal that Express-MaxDiff does a very good job in terms of recovering average parameters: Irrespective of the concrete experimental condition, the correlations between simulated and estimated mean parameters are constantly close to 1. Figure 5 illustrates the strengths of these relationships for a random experimental condition.

[5] As suggested in e.g. Silver, Dunlap 1987, the average is calculated after transforming the individual correlations to Fisher's Z values.

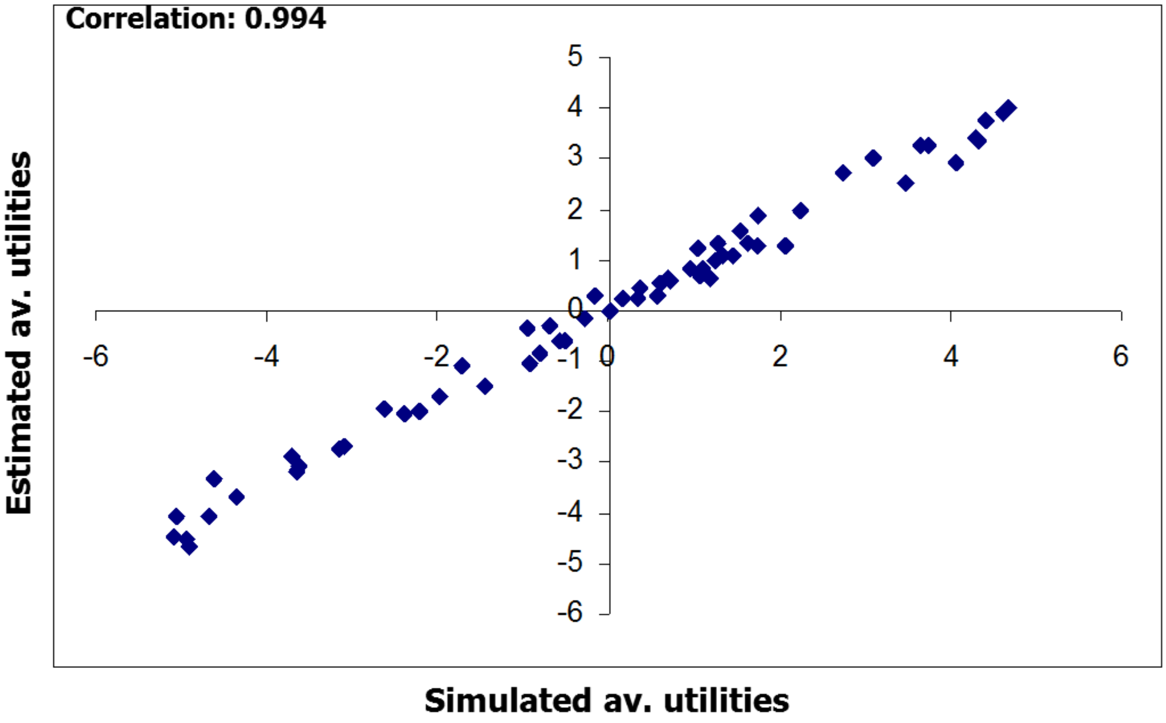
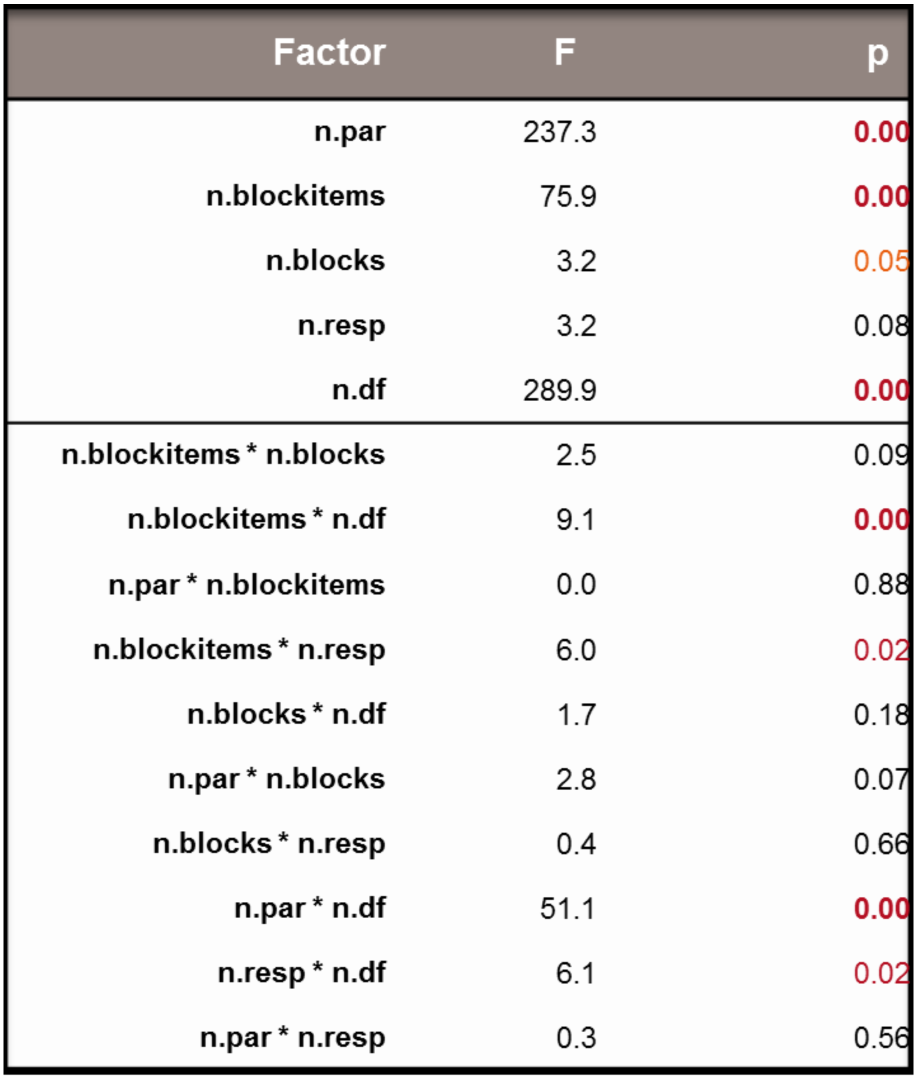


Figure 5
Example of Estimated versus True Mean Parameters

While this outcome regarding the recovery of aggregate preference structures is indeed convincing, the crucial question is, whether the Express-MaxDiff approach also leads to satisfactory results in terms of *individual* parameter recovery. As explained when introducing Express-MaxDiff, the main challenge in this respect is the fact that respondents see and evaluate only a part of the studied items. The remaining preference parameters have to be derived within the HB framework, i.e. based on population information, such as average preferences and the covariance matrix reflecting relationships between item-specific preferences.

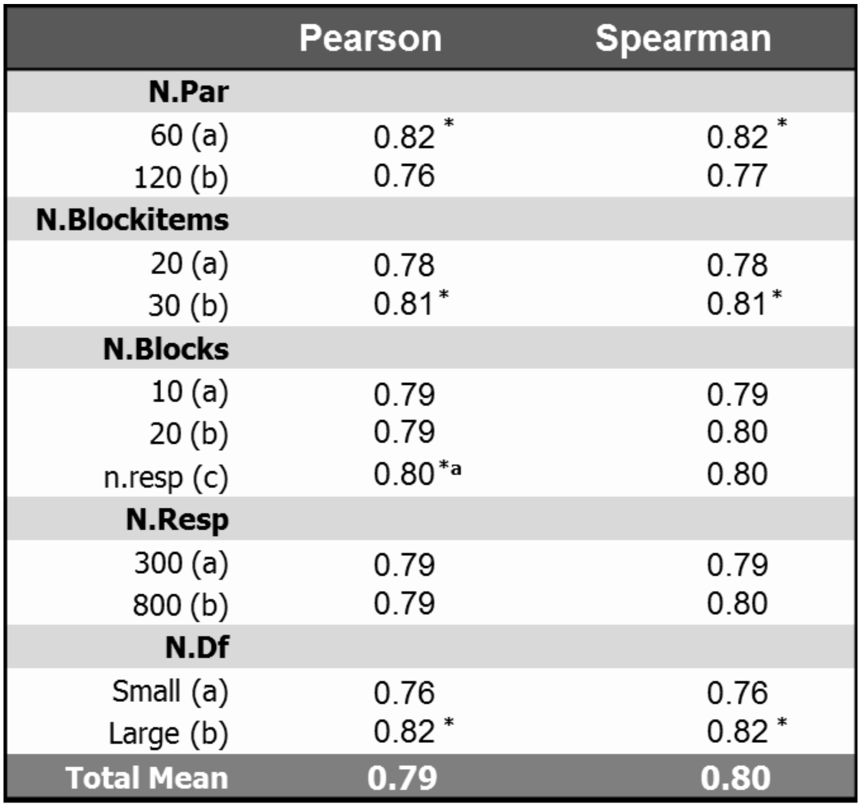
The results of the ANOVA that are presented in Table 2 give a first idea about the factors that most strongly influence the recovery of individual parameters: Obviously, the total number of parameters, the number of items in each design block, the number of design blocks and the number of prior degrees of freedom have a significant influence on individual parameter recovery. There are also some significant interactions that must be kept in mind when interpreting the results.



Factor	F	p
n.par	237.3	0.00
n.blockitems	75.9	0.00
n.blocks	3.2	0.05
n.resp	3.2	0.08
n.df	289.9	0.00
n.blockitems * n.blocks	2.5	0.09
n.blockitems * n.df	9.1	0.00
n.par * n.blockitems	0.0	0.88
n.blockitems * n.resp	6.0	0.02
n.blocks * n.df	1.7	0.18
n.par * n.blocks	2.8	0.07
n.blocks * n.resp	0.4	0.66
n.par * n.df	51.1	0.00
n.resp * n.df	6.1	0.02
n.par * n.resp	0.3	0.56

Table 2
ANOVA Table

More detailed insights can be obtained by analyzing the marginal means of the performance measure (i.e. the average correlation between true and estimated individual parameters), which can be found in Table 3. In addition to the Pearson correlation, which reflects the linear relationship between true and estimated parameters, the Spearman *rank* correlation is also taken into account in this analysis.



	Pearson	Spearman
N.Par		
60 (a)	0.82*	0.82*
120 (b)	0.76	0.77
N.Blockitems		
20 (a)	0.78	0.78
30 (b)	0.81*	0.81*
N.Blocks		
10 (a)	0.79	0.79
20 (b)	0.79	0.80
n.resp (c)	0.80*a	0.80
N.Resp		
300 (a)	0.79	0.79
800 (b)	0.79	0.80
N.Df		
Small (a)	0.76	0.76
Large (b)	0.82*	0.82*
Total Mean	**0.79**	**0.80**

Table 3
Marginal Means of Average Correlations between
True and Estimated Individual Parameters

The following conclusions can be drawn based on an analysis of these marginal means:

- The overall goodness of the individual parameter recovery is satisfactory: Taking into account the very challenging data situations – only between 16% (20 out of 120) and 50% (30 out of 60) of the items have actually been evaluated by each artificial respondent – overall correlations between true and estimated individual parameters of 0.79 (Pearson) and 0.80 (Spearman) are convincing.
- A lower total number of items and a larger number of items per block significantly[6] increase the quality of the individual parameter recovery. These observations match prior expectations: Fewer items result in fewer parameters to be estimated and of course *ceteris paribus* in a better estimation. More items per design block are equivalent to more items being actually seen and evaluated by respondents, i.e. more “true” individual preference information that does not have to be imputed.
- The number of design blocks does not seem to have a strong influence on the goodness of the estimation. Even the extreme case, in which each respondent gets an individual design block, does only lead to marginally superior estimates at best.

[6] Significant differences at the 0.05 level are reflected by a “*” that is attached to the superior mean.

- The sample size does not have a significant positive effect on the goodness of the estimation. While this might appear surprising at first sight, it matches observations made in several previous studies, such as Wirth (2010, 2011).
- In contrast, a very strong positive effect can be observed when looking at the number of prior degrees of freedom. Obviously, estimation settings characterized by a larger number of prior degrees of freedom are clearly beneficial when estimating parameters based on Express-MaxDiff designs. This makes intuitive sense, as *ceteris paribus* more prior degrees of freedom lead to a stronger Bayesian shrinkage. In other words: the larger the number of degrees of freedom, the more influence the population information on the estimation of individual parameters. Given the crucial role of population information in the Express MaxDiff approach, it makes sense that the estimation gets better in these cases.

All in all, the Monte-Carlo simulation study reveals a convincing performance of Express-MaxDiff: On the aggregate level, parameters are recovered almost perfectly, and the recovery of full individual parameter vectors is also satisfactory – especially when taking into account the very challenging investigated data situations. Furthermore, the following *key success factors* can be identified:

- the number of prior degrees of freedom (c.p. a higher number leads to a better recovery of individual parameters),
- the number of items per design block (the more items the respondents actually evaluate, the better the recovery of the full parameter vectors), and
- the total number of parameters (of course, the fewer parameters are to be estimated, the better parameter recovery is).

However, as mentioned before, a convincing performance in Monte-Carlo simulation studies is necessary, but not sufficient, for superiority based on real data. In order to validate the simulation-based findings, we have conducted an empirical comparison study, which is described in the next section.

Empirical Comparison Study

The empirical study was designed in order to enable a fair comparison of the introduced Express-MaxDiff approach and two alternative approaches to MaxDiff for large item sets, namely Augmented MaxDiff and an approach that we call Sparse MaxDiff:

> As briefly described in the introductory section of this paper, the main idea behind *Augmented MaxDiff* is an enrichment of preference information derived from a sparse MaxDiff exercise with preference information derived from an additional Q-Sort task. Many details about this approach can be found in Hendrix and Drucker (2008).
>
> In contrast, *Sparse MaxDiff* can be seen as a specific variant of Express-MaxDiff: While Express MaxDiff is based on the idea of letting respondents evaluate *a subset of all items* more *thoroughly*, respondents have to evaluate *all* items in Sparse MaxDiff, but by far not as thoroughly. For example, an Express-MaxDiff design might result in respondents seeing and evaluating each item in their respective design block at least three times and the remaining items not at all. However, in a Sparse MaxDiff design, respondents see and evaluate all items, but usually just once.

STUDY DESIGN

The three different MaxDiff approaches were tested in three equally structured samples of planned n=500 in the US. Each sample (= “cell”) was representative for the US population with regard to gender and age. The study dealt with the individual importance of 60 personal values taken from GfK Roper Consumer Styles. Judged by the positive comments, respondent involvement was unusually high. A summary of the study design can be seen in Table 4.

	Study characteristics
Population	US population
Sample size	N=1498
Sample quota	Gender x Age (M/F x Age 18-29; 30-39; 40-49; 50+)
Items	60 personal values (GfK Roper Styles)
Assignment	Random assignment within strata to one of three cells

Table 4
Design of Empirical Study

In order to be able to compare the performance of the three MaxDiff approaches, three ranking questions including five items each were added as holdout tasks at the end of each questionnaire. Furthermore, qualitative questions were asked to get an idea about the respondents’ evaluation of the interview process.

Each respondent was randomly assigned to one of the three cells, i.e. to one approach. The specific study settings in these cells are displayed in Table 5 and explained in more detail in the following.

MaxDiff approaches	# of items (total/ subset)	MD-tasks per respondent	# of items per MD-set	# of times each item is displayed per design version	N (resp.)
Express MD	60/20	12	5	0 or 3 (block design)	500
Augmented MD	60/60	12+Q-Sort (3 “buckets”)	5	1+Q-Sort (3 “buckets“)	499
Sparse MD	60/60	12	5	1	499

Table 5
Design Settings in Different Cells

Cell 1—Express MaxDiff:

Each respondent had to evaluate a subset of 20 (out of the 60) items. The design blocks were individualized, i.e. we generated as many design blocks as there were respondents in the sample. Therefore, for each design block, only one MaxDiff questionnaire version was created. Each version contained 12 MaxDiff tasks, with 5 items per task. Using Sawtooth Software’s MaxDiff Designer, these settings mean that each item has been displayed three times on average per questionnaire version.

Cell 2—Augmented MD:

In the Q-Sort section, respondents had to assign the 60 items to three buckets according to their importance. We implemented this by using drag & drop questions. First, the 20 most important items (i.e. personal values) had to be chosen this way, then the 20 least important items. Obviously, this procedure resulted in three “buckets”, each containing 20 items. While after the Q-Sort task still nothing is known about the relative importance of items *within* one specific bucket, it is the implicit information about the relative importance of items in *different* buckets that is used during the estimation (see Hendrix, Drucker 2008).

The Q-Sort section was followed by 12 standard MaxDiff questions. Like in the Express-MaxDiff cell, each MaxDiff task contained five items. However, in contrast to Express-MaxDiff, the entire list of 60 items is seen and evaluated by each respondent. These settings result in each item being exposed on average only once per questionnaire.

Cell 3—Sparse MaxDiff:

As described above, Sparse MaxDiff is characterized by the fact that all items are evaluated by each respondent just once. Therefore, the same design settings as in the MaxDiff section of the Augmented MaxDiff cell could be used. However, no additional tasks, such as Q-Sort questions, were added.

RESULTS

The comparison of the three approaches is based on the following criteria:

- Respondents' stated evaluations of the questionnaires resulting from the three approaches,
- interview duration,
- correlations of average importance scores and
- predictive validity, assessed through holdout evaluations.

The average *subjective evaluations* of the three questionnaires with regard to different dimensions can be found in Table 6. It can be seen that all approaches are evaluated very similarly. The only statement for which significant[7], although not particularly strong, differences in the average ratings can be observed, relates to the perceived ability to express one's opinion.

	Respondents' Ratings on a 1-7 scale (1: strongly disagree, 7: strongly agree)				
	… was enjoyable	… was confusing	… was easy	… made me feel like Clicking Through	… allowed me to express my opinion
Express MaxDiff (1)	5.59	2.41	5.53	2.06	5.78
Augmented MaxDiff (2)	5.68	2.47	5.54	2.05	5.97^{1}
Sparse MaxDiff (3)	5.69	2.35	5.53	2.07	5.95^{1}

Table 6
Average Stated Evaluations of Approaches

In contrast, substantial differences can be observed in terms of *interview duration*. Due to the additional Q-Sort section, Augmented MaxDiff leads to a substantially longer interview duration compared to Express and Sparse MaxDiff – in our study, it took respondents on average about 3-5 minutes longer to complete an Augmented MaxDiff study than to complete a Sparse

[7] In all tables, superscripts denote significant differences. For example, a "1" means that the marked value is significantly higher than the value of the first approach.

MaxDiff or an Express MaxDiff study (see Table 7). Practitioners should take this into consideration when planning to use Augmented MaxDiff.

When looking at the *correlations of the average importance scores* between the 3 approaches (see scatter plots in the appendix), it can be found that the average results of all approaches correlate relatively closely (all correlation coefficients are greater than 0.90). This indicates that the average results seem to agree at a certain level, although the scatter plots show that the average results of Augmented MaxDiff still seem to be a bit different from the results of the two other approaches – maybe because of the integration of different sources of preference information.

	Mean interview duration (in min:s)	Median interview duration (in min:s)
Express MaxDiff	18:57	12:00
Augmented MaxDiff	22:17	15:00
Sparse MaxDiff	17:33	13:00

Table 7
Mean Interview Durations

When assessing *predictive validity*, i.e. the answers to the three holdout ranking tasks, we follow Hendrix' and Drucker's (2008) approach: We evaluate the percentage of correctly predicted best and worst answers, the average number of items that are predicted correctly and Spearman's rho between actual and predicted ranks.

It can be observed (see Table 8) that all three approaches perform satisfactory in terms of predictive validity and that the comparable figures are similar to the ones observed by Hendrix and Drucker (2008). All in all, Augmented MaxDiff seems to be slightly superior regarding all measures, but surprisingly the performance of Sparse MaxDiff – which based on our settings (see Table 5) is basically Augmented MaxDiff without the Q-Sort section – comes very close.

	Predictive Validity			
MaxDiff Models	Correct Bests (%)	Correct Worsts (%)	Avg. #items correct	Avg. Spearman Rho
Express MaxDiff	61.7%	54.1%	2.25	0.633
Augmented MaxDiff	66.3%	62.8%[1]	2.52[1]	0.702[1]
Sparse MaxDiff	62.7%	61.7%[1]	2.43[1]	0.691[1]

Table 8
Predictive Validity of the Compared Approaches

Express-MaxDiff is doing well in terms of predicting the best alternatives in the holdout tasks: The hit rate is not significantly lower than the hit rate of the other approaches. However, when it comes to correctly predicting the worst alternatives, Express-MaxDiff is significantly inferior, although still at a level that is much higher than a naïve prediction. This flaw of Express MaxDiff very likely also causes the inferiority in terms of the other two measures. The reason(s) for Express-MaxDiff performing well with regard to predicting rather important items, but weaker when it comes to predicting rather unimportant items are unclear and require further testing.

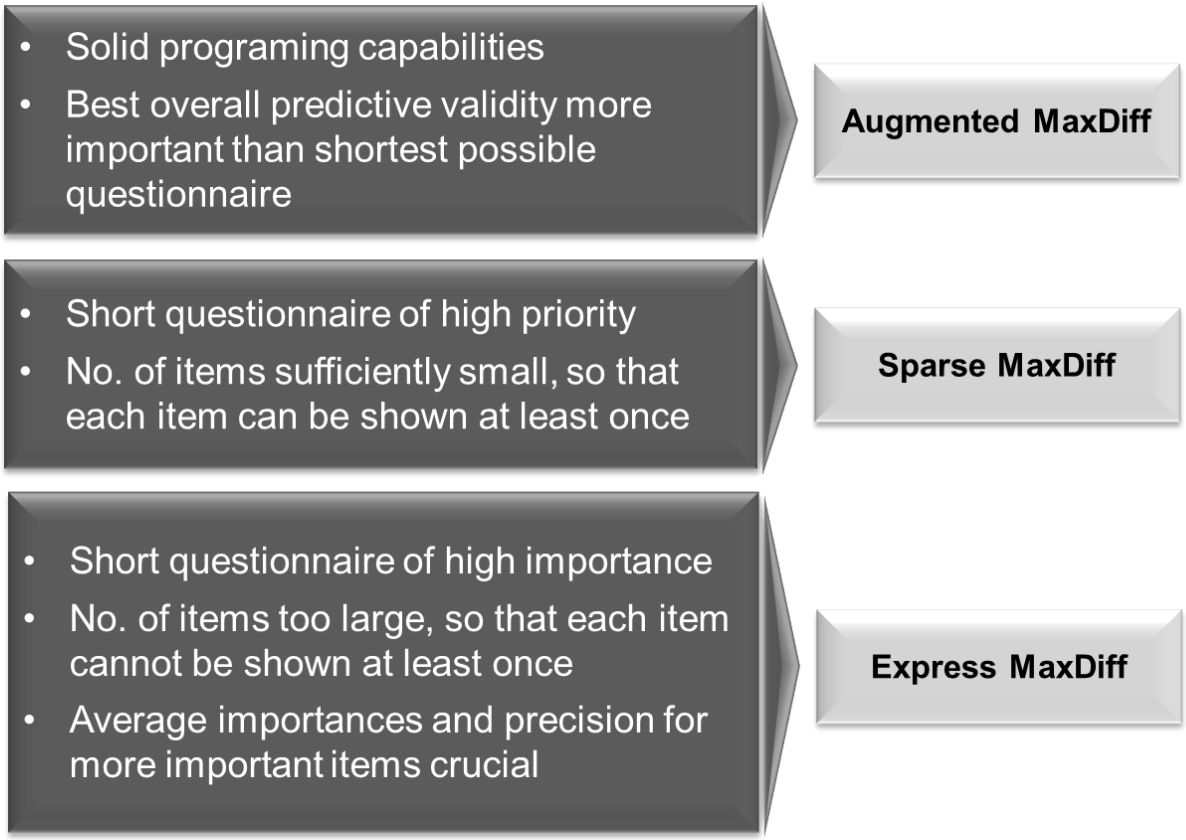
All in all, based on our empirical study one can conclude that a combination of Q-Sort and a sparse MaxDiff design like in the Augmented MaxDiff approach seems to be most promising when it comes to estimating individual parameters in MaxDiff studies dealing with large numbers of items. However, this comes at the price of a significantly longer questionnaire. For that reason, Sparse MaxDiff and Express MaxDiff are attractive alternatives: Both perform very well in terms of predicting more important items and Sparse MaxDiff does so even in terms of predicting less important items.

Summary and Recommendations

Augmented MaxDiff, introduced by Hendrix and Drucker (2008) is a powerful approach to MaxDiff studies with very many items. Our empirical comparison study has confirmed that in a convincing manner. However, it has also been shown that the Q-Sort section required for Augmented MaxDiff adds significant time requirements. Therefore, alternative approaches characterized by shorter questionnaires may be valuable alternatives in some situations.

Both our Monte Carlo simulation study and our empirical comparison study have shown that *Express-MaxDiff* is able to deal with the challenges of MaxDiff studies involving many items. Although respondents see and evaluate only a subset of all items, average parameters are recovered almost perfectly (see Figure 5) and the performance regarding individual parameter recovery and empirical predictive validity is satisfactory, too (see Table 3 and Table 8). Nevertheless, a specific variant of Express MaxDiff that we call *Sparse MaxDiff* also performs very well in the empirical comparison and might be the method of choice in situations where Sparse MaxDiff designs are feasible, i.e. when the number of items is sufficiently small to ensure that each single item is evaluated at least once by each respondent.

Our final recommendations are summarized in Figure 6.



- Solid programing capabilities
- Best overall predictive validity more important than shortest possible questionnaire

Augmented MaxDiff

- Short questionnaire of high priority
- No. of items sufficiently small, so that each item can be shown at least once

Sparse MaxDiff

- Short questionnaire of high importance
- No. of items too large, so that each item cannot be shown at least once
- Average importances and precision for more important items crucial

Express MaxDiff

Figure 6
Final Recommendation

References

Allenby, Greg M.; Rossi, Peter E. (1999): Marketing Models of Consumer Heterogeneity. Journal of Econometrics 89(1/2), 57–78.

Allenby, Greg M.; Bakken, David G.; Rossi, Peter E. (2004): The HB Revolution: How Bayesian Methods Have Changed the Face of Marketing Research. Marketing Research 16(2), 20–25.

Allenby, Greg M.; Rossi, Peter E. (2003): Perspectives Based on 10 Years of HB in Marketing Research. Sawtooth Software Inc. Sequim, WA. Sawtooth Software Research Paper Series. http://www.sawtoothsoftware.com/download/techpap/allenby.pdf.

Bass, Frank M.; Wilkie, William L. (1973): A Comparative Analysis of Attitudinal Predictions of Brand Preference, Journal of Marketing Research, 10, 262-269.

Baumgartner, Hans; Steenkamp, Jan-Benedict E. M. (2001): Response Styles in Marketing Research: A Cross-National Investigation. In: Journal of Marketing Research, 38, 143–156.

Beckwith, Neil E.; Lehmann Donald (1973): The Importance of Differential Weights in Multi-Attribute Models of Consumer Attitude, Journal of Marketing Research, 10, 141-145.

Chrzan, Keith (2008): Measuring Attribute Importance: Stated and Derived Methods. Workshop at the SKIM European Conference 2008, Barcelona.

Chrzan, Keith; Golovashkina, Natalia (2006): An Empirical Test of Six Stated Importance Measures. In: International Journal of Market Research, 48, 717–740.

Cohen, Steven H. (2003): Maximum Difference Scaling: Improved Measures of Importance and Preference for Segmentation. 2003 Sawtooth Software Conference Proceedings. Sequim, WA, 61–74.

Cohen, Steven H.; Orme, Bryan K. (2004): What's Your Preference? Asking Survey Respondents About Their Preferences Creates New Scaling Dimensions. In: Marketing Research, 16, 33–37.

Finn, Adam; Louviere, Jordan J. (1992): Determining the Appropriate Response to Evidence of Public Concern. The Case of Food Safety. In: Journal of Public Policy and Marketing, 11, 12–25.

Hendrix, Phil; Drucker, Stuart (2008): Alternative Approaches to Maxdiff With Large Sets of Disparate Items – Augmented and Tailored Maxdiff. 2007 Sawtooth Software Conference Proceedings, 169-188.

Marley, Anthony A. J.; Louviere, Jordan J. (2005): Some Probabilistic Models of Best, Worst, and Best-Worst Choices. In: Journal of Mathematical Psychology 49, 464–480.

Rässler, Susanne; Koller, Florian; Mäenpää, Christine (2002): A Split Questionnaire Survey Design applied to German Media and Consumer Surveys Discussion paper 42a/2002 Online: http://statistik.wiso.uni-erlangen.de/forschung/d0042b.pdf

Rossi, Peter E.; Gilula, Zvi; Allenby, Greg M. (2001): Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach. In: Journal of the American Statistical Association, 96, 20–31.

Silver, N. Clayton; Dunlap, William P. (1987): Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used. In: Journal of Applied Psychology 72(2), 146–148.

Wilkie, William L; Pessemier Edgar A. (1973): Issues in Marketing's Use of Multi-Attribute Attitude Models, Journal of Marketing Research, 10, 428-441.

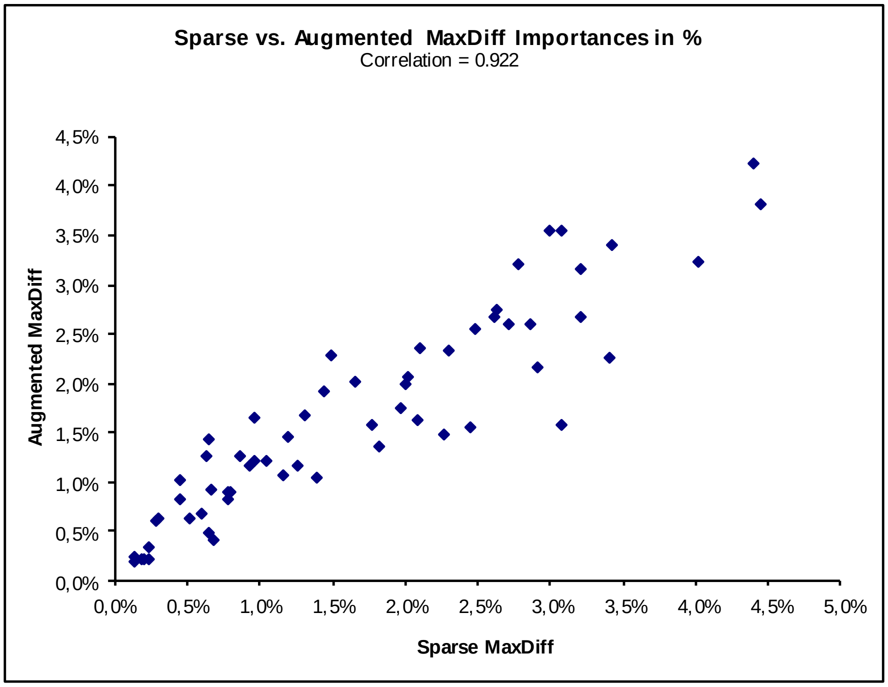
Wirth, Ralph (2008): Good - Better - Best-Worst? New Empirical Findings on Best-Worst Scaling. 2008 International Meeting of GfK Statisticians and Methodologists (IMSM), Madrid.

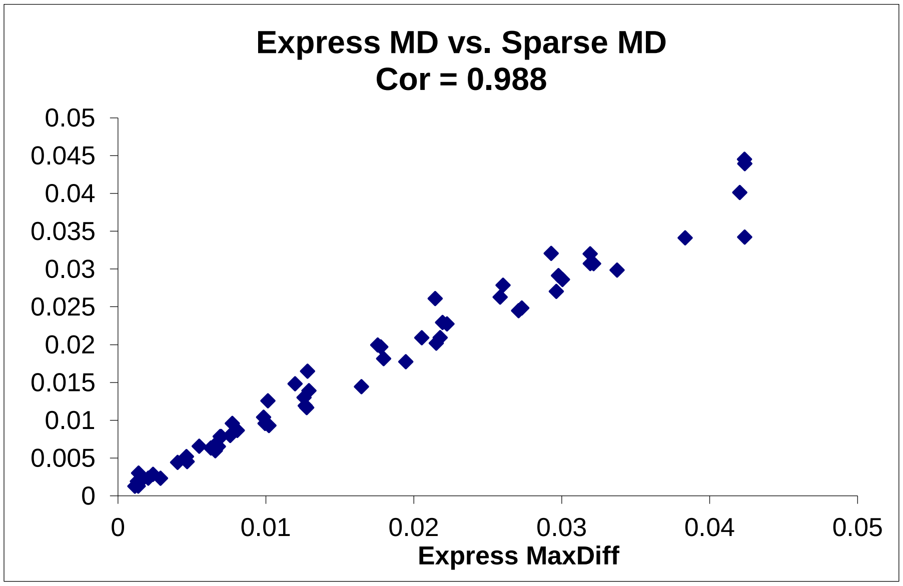
Wirth, Ralph (2010): Best-Worst Choice-Based Conjoint-Analyse. Eine neue Variante der wahlbasierten Conjoint-Analyse. Marburg: Tectum-Verlag.

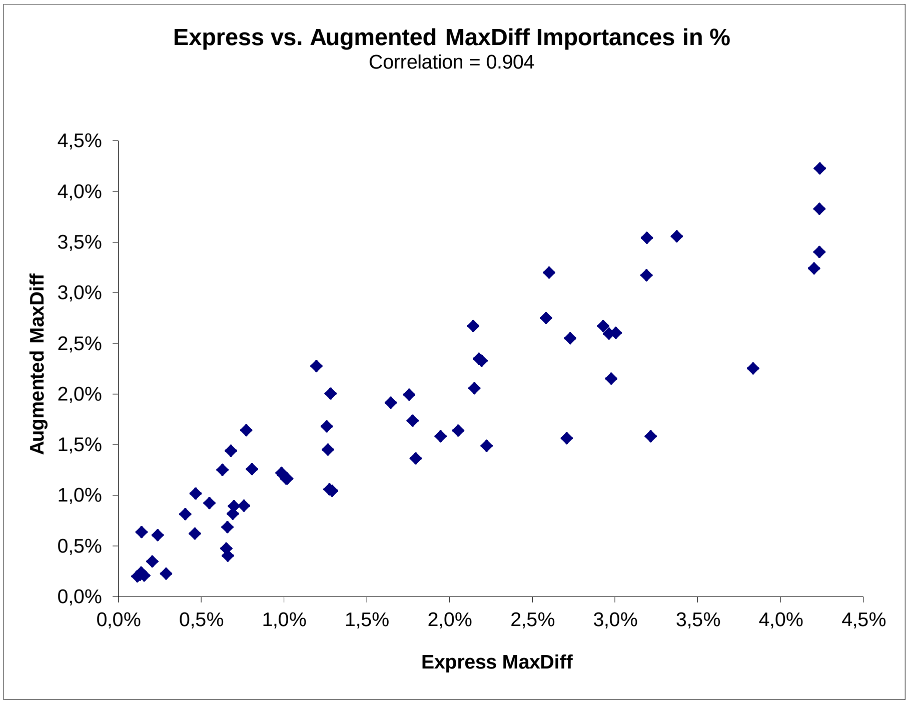
Wirth, Ralph (2011): HB-CBC, HB-Best-Worst-CBC or No HB At All? 2010 Sawtooth Software Conference Proceedings, 321-354.

Appendix

Correlations and scatter plots of the average importance scores between the 3 approaches




Express MD vs. Sparse MD
Cor = 0.988
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.01
0.02
0.03
0.04
0.05
Express MaxDiff


Express vs. Augmented MaxDiff Importances in %
Correlation = 0.904
Augmented MaxDiff
4,5%
4,0%
3,5%
3,0%
2,5%
2,0%
1,5%
1,0%
0,5%
0,0%
0,0%
0,5%
1,0%
1,5%
2,0%
2,5%
3,0%
3,5%
4,0%
4,5%
Express MaxDiff

WHAT'S IN A LABEL?
BUSINESS VALUE OF HARD VERSUS SOFT CLUSTER ENSEMBLES

NICOLE HUYGHE
ANITA PRINZIE
SOLUTIONS-2

SUMMARY

Cluster ensembles improve segmentations. Hard cluster ensembles, combining segment labels from multiple cluster solutions, are popular. Combining soft segment membership probabilities is as simple but surprisingly less common. We assess the added business value of soft cluster ensembles on three industry projects and on simulated data. The industry experiments provide mixed evidence. The simulation study however reveals that hard cluster ensembles provide more distinct segments and more balanced segments than soft cluster ensembles do. As segment heterogeneity and segment size are key to most industry segmentations, hard cluster ensembles are recommended.

CLUSTER ENSEMBLES ANSWER THE WHO AND WHY

The average consumer does not exist, nor does the successful average product. That's why smart managers tailor their products and marketing communication to the different needs, attitudes, values, etc. of customer segments. It is assumed that different buying behavior and response to marketing stem from customers' differences in needs, beliefs, attitudes, values, socio-cultural environment and so on. Hence, smart managers segment the market into clusters of customers and employ segment-specific marketing strategies. Within a segment customers have similar needs and attitudes, hence are likely to behave and respond similarly. Ideally, a marketing manager wants segments to differ on needs and attitudes as well as on more tangible attributes such as behavior and socio-demographics. If so, he will know why customers in a segment behave as they do and he will know who these segment members are, so he can target them. Answering the 'who' and 'why' are key to successful segment-specific marketing strategies.

A standard segmentation method such as k-means or hierarchical clustering will struggle to find segments which differ both on the "who" and "why" (i.e. attitudes & needs *and* behavior & socio-demographics). Suppose a manager conducts a survey to identify customer segments. The survey will contain sets of questions or themes representing the dimensions on which he would like the segments to differ (Figure 1). Using traditional segmentation methods, all themes from the survey will be used as segmentation drivers. Unfortunately, the resulting segments are likely to be dominated by only a few themes, e.g. theme 1 (brand attitude) and theme 3 (needs); unlikely to answer both the "who" and "why". This is where cluster ensembles come in.

Instead of running a segmentation analysis on all themes simultaneously, feature-distributed cluster ensembles (Strehl and Gosh, 2002) make use of a separate traditional segmentation on each survey theme. Subsequently, these theme-specific segmentation solutions are used as input to yet another segmentation. The latter acts like a meta-segmentation. By segmenting theme-specific segmentations the cluster ensemble will find segments which differ on all themes rather than just a few. That's why cluster ensembles answer the "who" and the "why".

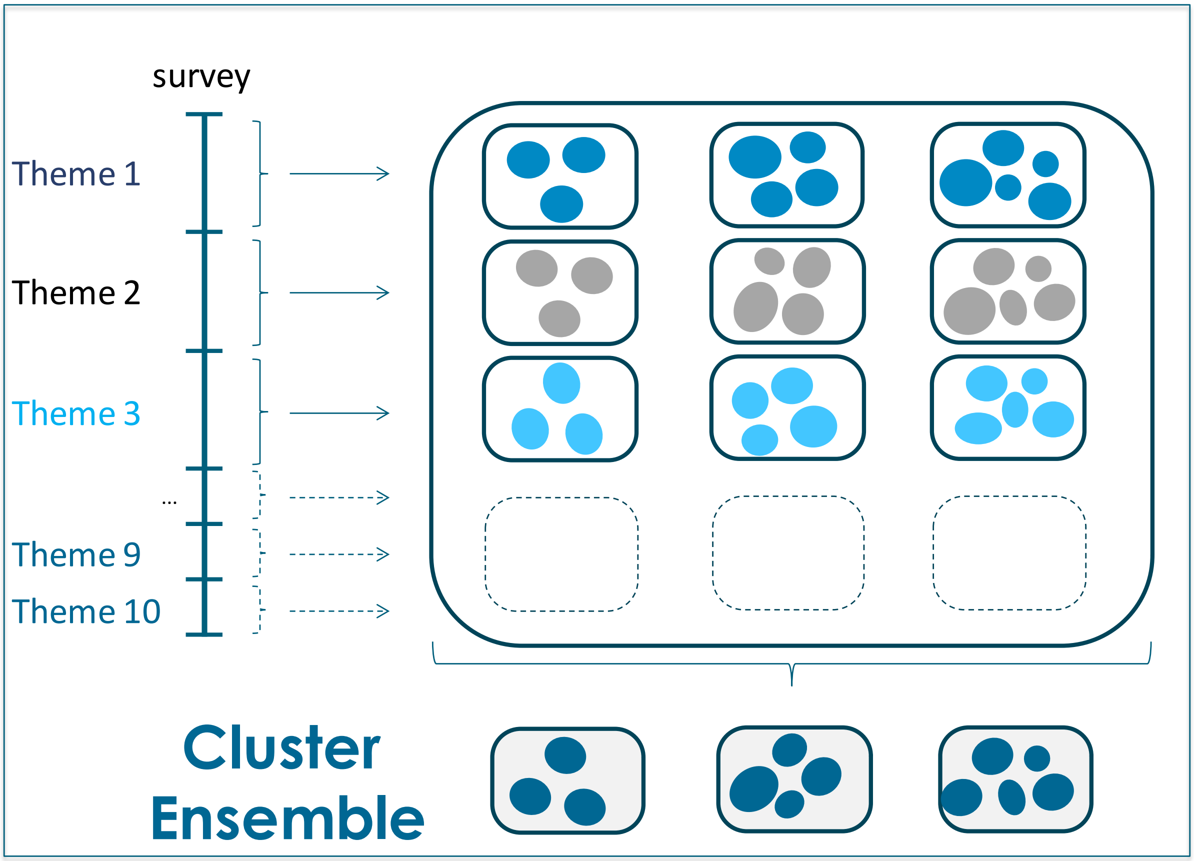


Figure 1: Cluster Ensembles for detecting segments differing on multiple survey themes

Soft Cluster Ensembles Use More Granular Information than Hard Cluster Ensembles

Cluster ensembles improve industry segmentations by answering both the "who" and "why". Moreover, cluster ensembles are a prominent method for improving robustness, stability and accuracy of the segmentation (Fred and Jain, 2002). There are different types of cluster ensembles, among which are the so called "hard" and "soft" cluster ensembles.

The difference between a "hard" and a "soft" cluster ensemble involves the level of granularity of the segmentation solutions on which the meta-clustering is done. "Hard" cluster ensembles combine segment labels from multiple cluster solutions into a meta-segmentation. For example, Sawtooth Software's CCEA (Convergent Cluster & Ensemble Analysis) creates a cluster ensemble by estimating a k-means clustering on the class labels of all cluster solutions. "Soft" cluster ensembles adopt a more granular approach than "hard" ensembles. Instead of combining segment labels from the multiple clusterings, the cluster membership probabilities (more granular than hard segment labels) are used as input to the meta-segmentation. These cluster ensembles are referred to as "soft" cluster ensembles as they use "soft" evidence indicating the degree of cluster membership.

In Figure 2 we display the level of granularity of the information used by a hard cluster ensemble (left) and by a soft cluster ensemble (right). A hard cluster ensemble will see that the lady has blue eyes (lady to the left) and feeds this high-level information into the meta-segmentation. On the other hand, a soft cluster ensemble uses more granular information. It will use a magnifying glass and will see that the lady actually has blue eyes with black eyeliner (lady to the right). This more granular information is used by the meta-clustering algorithm.

Figure 2: More granular information used by soft cluster ensembles

Business Value of Hard vs. Soft Cluster Ensembles

"Hard" cluster ensembles, combining segment labels, are far more popular than soft cluster ensembles, which combine cluster membership probabilities. This is somewhat surprising as "soft" cluster membership data are easily obtained from latent class or fuzzy clustering methods. Besides, combining "soft" clusterings is not more complex than combining "hard" clusterings (Topchy, Jain and Punch, 2004). Furthermore, we assume that using more granular information such as segment membership probabilities could be beneficial to the segmentation. For example, when the underlying segmentation is not very confident in the segment allocation (e.g. 35% probability to belong to segment 1 in a 3 segment solution), using the cluster membership probabilities (P1=33%, P2=32%, P3=35%) is more informative (revealing the fuzziness) than using the segment label (3) and therefore might benefit the segmentation.

This paper assesses the business value of "hard" versus "soft" cluster ensembles. The business value is assessed on three industry data sets as well as on simulated data. The segmentation solution from a "hard" or "soft' cluster ensemble is evaluated with respect to stability, integrity (homogeneity and heterogeneity), accuracy and size. The next subsections explain each of these criteria in further detail. The first paragraph always explains the rationale behind each criterion. Readers with less interest in the technical steps followed in each section can skip the more technical paragraphs.

Stability

A good segmentation should have stable segments. Segments are less stable if a large percentage of observation pairs no longer belong together when adding an extra segment. In Figure 3, the two blue persons (solid oval in C) stay in the same segment after adding a fourth segment, but the two green persons (dashed oval in C) no longer belong to the same segment after adding one extra segment.

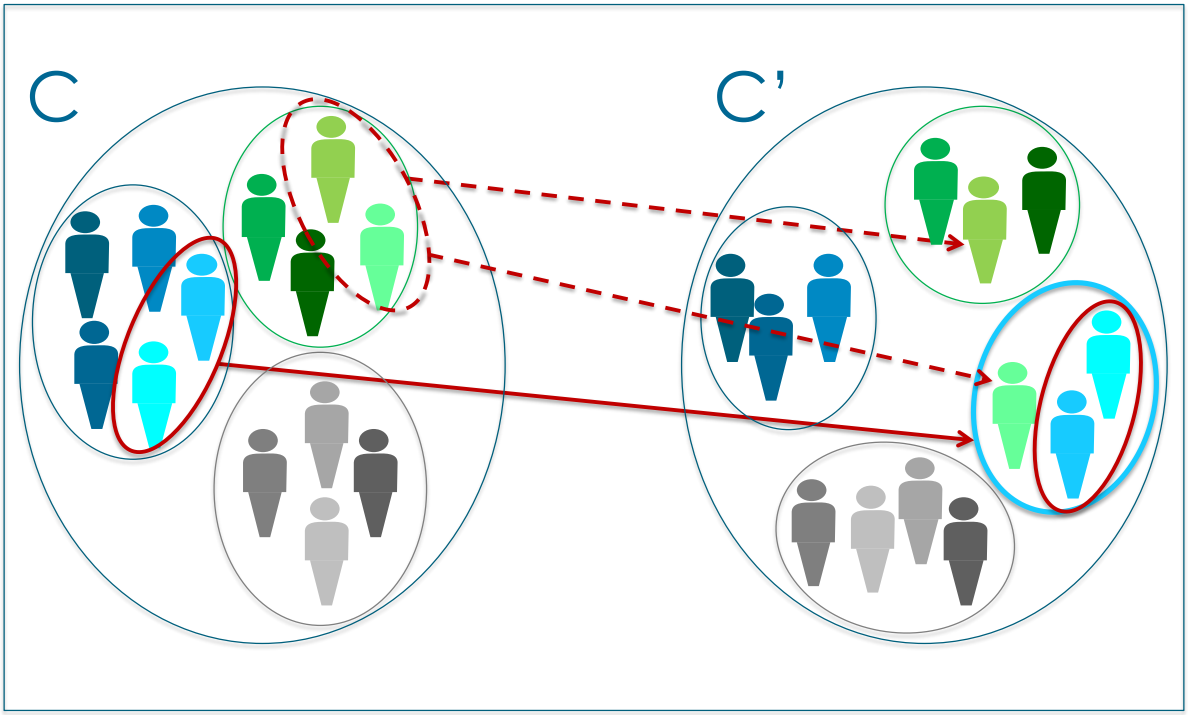


Figure 3: stability intuition

The *similarity index* (Lange et al., 2004) is used as a measure of segment stability. It indicates the percentage of pairs of observations that belong to the same cluster in both clustering C and clustering C' (Equations 1 and 2). In our case, C' refers to segmentation with one additional segment added to the original segmentation C.

$$Similarity\ Index = \frac{dpCC'}{\sqrt{dpCC \times dpC'C'}} \qquad (1)$$

$$dpCC' = \sum_{1}^{i,j} C[i,j] \times C'[i,j] \qquad (2)$$

with

$C[i,j] =$ 1 if objects i and j belong to the same segment in cluster solution C
$=$ 0 if objects i and j belong to a different segment in cluster solution C

To calculate the stability in this research, we take the average of the similarity indices for a 3 to 4 cluster solution, a 4 to 5 cluster solution and a 5 to 6 cluster solution. We programmed the similarity index using SPSS PASW Statistics 19 software.

Integrity: maximizing heterogeneity and homogeneity

The hard and soft cluster ensembles are also evaluated with respect to cluster integrity.

The segmentation's integrity is high when people within a segment are similar (high homogeneity) and people from different segments are highly distinct (high heterogeneity). Segment homogeneity and heterogeneity are key to any business segmentation. A manager wants customers from different segments to have a different profile so they will behave and react differently. Likewise, he wants the customers within a segment to be as similar as possible so they will behave and react similarly.

The segment *heterogeneity* is measured by the *total separation* metric (Halkidi, Vazirgiannis and Batistakis, 2000) which is based on the distance between cluster centers. In Equation 3 D_{max} is the maximum distance between cluster centers, D_{min} is the minimum distance between cluster centers.

The lower the total separation, the more heterogeneous the segments are.

$$Total\ Separation = \frac{D_{max}}{D_{min}} \sum_{k=1}^{n_c} \left(\sum_{z=1}^{n_c} \| v_k - v_z \| \right)^{-1} \qquad (3)$$

with

D_{max}	$\max(\|v_i - v_j\|)\ \forall i, j \in \{1,2,3,\dots,n_c\}$
D_{min}	$\min(\|v_i - v_j\|)\ \forall i, j \in \{1,2,3,\dots,n_c\}$
n_c	number of clusters
v_k	cluster center of cluster k
v_z	cluster of cluster z, with $z \neq k$
$\|x_k\|\|$	$(x_k^T x_k)^{1/2}$

The cluster *homogeneity* assesses how similar subjects are within a segment. It is measured by the segment *scatter* (Halkidi, Vazirgiannis and Batistakis, 2000) which starts from the ratio of the cluster variance to the variance in the total sample and averages these ratios over all segments. The lower the scatter the more homogeneous the segments are.

$$Scatter = \frac{1}{n_c}\sum_{i=1}^{n_c}\frac{\|\sigma(v_i)\|}{\|\sigma(x)\|} \qquad (4)$$

with

n_c	number of clusters
$\sigma(v_i)$	variance of cluster i
$\sigma(x)$	variance of the dataset
$\|x_k\||$	$(x_k^T x_k)^{1/2}$

We use the CLV R package implementation of the total separation and scatter using respectively the clv.dis and clv.scatt functions (Nieweglowski, 2007).

Accuracy

For a manager the segments should be predictable so he can implement the segmentation and employ segment-specific marketing strategies. The manager should be able to allocate a customer to one of the identified segments so he can tailor his marketing actions accordingly.

The predictability of the segments is expressed by the adjusted Rand Index (Hubert and Arabie, 1985). It expresses the level of agreement between the predicted segment and the real segment correcting for the expected level of agreement.

The adjusted Rand Index starts from a confusion table. The higher the adjusted Rand Index the higher the accuracy.

$$Adjusted\ Rand\ Index = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \frac{\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}}{\binom{n}{2}}}{\frac{\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}}{2} - \frac{\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}}{\binom{n}{2}}} \qquad (5)$$

with

n_{ij}	number of objects that are both in predicted real segment ui and cluster vj
$n_{i.}$	number of objects that are in real segment u_i
$n_{.j}$	number of objects that are in predicted segment u_j

The segment prediction accuracy is assessed for the industry data sets only. On the industry data sets, we either use a multinomial logit algorithm or a discriminant analysis to predict the segments using the survey questions. On the simulated data, we cannot assess the prediction accuracy as we do not have any predictors (survey questions) to predict the real segment.

Size

Employing segment-specific marketing strategies typically does not pay off for very small segments. On the other hand, when one of the segments is very large with the other segments being very small, the quality of the segmentation might be questionable. Therefore, the hard and soft cluster ensembles are evaluated with respect to segment size.

In a perfectly balanced segmentation segments are equal in size. In that case the segment distribution is uniform. To measure how well balanced the segments are with respect to size we calculate the uniformity deviation. This is the extent to which the segment size deviates from equal segment size or the expected segment size given a uniform segment distribution.

We calculate this uniformity deviation for each segment, take the absolute form of this deviation and average these absolute deviations over the segments. The calculation of the average absolute uniformity deviation can be explained by a simple example. Suppose you have a three cluster solution with segment proportions s1=25%, segment proportion s2=30% and segment proportion s3=45%. The absolute uniformity deviation for each of the three segments is respectively |25%-33%|=8, |30%-33%|=3 and |45%-33%|=12. The average absolute deviation for this 3 cluster solution is then |(8+3+12)/3| or 7.6. Large average absolute uniformity deviation indicates that the clustering has segments that are very unequal in size, so potentially having very small or very large segments. In the following text, we will refer to uniformity deviation as shorthand notation for average absolute uniformity deviation.

$$Uniformity\ Deviation = \sum_{c=1}^{C} |prop_c - prop_U| \tag{6}$$

Industry Experiments

Business value of hard versus soft cluster ensembles using multiple Key Performance Indices

We assess the business value of hard and soft cluster ensembles on three industry datasets: rheumatism (N=400), osteoporosis (N=450) and customer software journey (N=206). The rheumatism and osteoporosis datasets segment either patients or physicians. In the customer software journey project B2B customers are segmented with respect to the factors that are key to a successful software adoption process.

We use the previously discussed evaluation criteria: stability, integrity (homogeneity and heterogeneity), accuracy and size. The integrity is measured by calculating the scatter (homogeneity) and total separation (separation) on multiple Key Performance Indicators (KPIs). The KPIs are project specific and vary from likelihood to prescribe a particular drug, belief that drug X is the logical next step in therapy to likelihood to recommend new software.

We calculate the scatter and total separation for each KPI separately. We refrain from calculating a single scatter metric or a single total separation metric including all KPIs. After all, such an aggregated metric fails to provide a detailed insight on the particular KPIs the segments are homogeneous on and the particular KPIs the segments are well-differentiating on. When deciding which segmentation to go for from a management perspective, it is useful to have a more detailed insight into which KPIs the segments are well-differentiated on and which KPIs the segments are homogeneous on.

To summarize the multiple scatter metrics calculated for each KPI separately, we calculate the number of wins across the KPIs rather than taking an average scatter. Likewise, to summarize the total separation, we calculate the number of wins across the KPIs. A win is notated when the scatter/total separation for the soft cluster ensemble is better than for the hard cluster ensemble on the particular KPI.

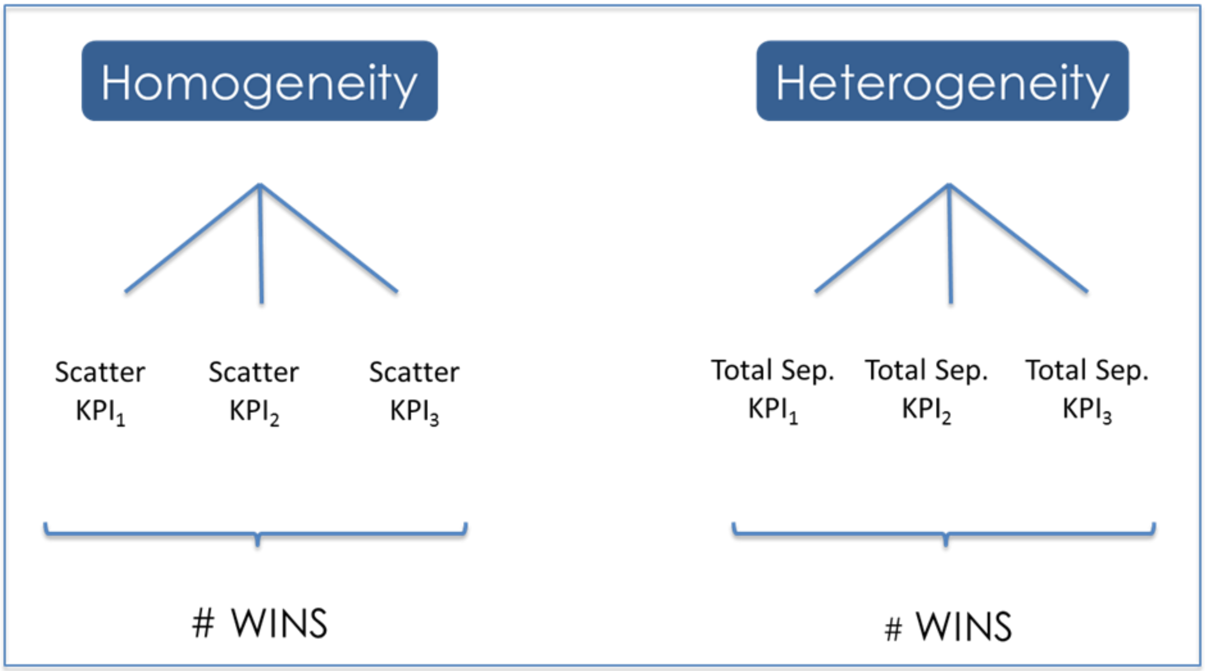


Figure 4: Scatter and total separation calculation on multiple KPIs

Mixed Evidence

For each of the three datasets the survey is divided into themes tapping into the "who" and "why" of patients, physicians or customers. On each of these themes a latent class clustering algorithm (Latent Gold 4.5) creates theme-specific segments. The segment labels or segment probabilities of these theme-specific segmentations are used as input to another clustering algorithm; the so called meta-clustering algorithm. To make the comparison between hard and soft cluster ensembles independent of the meta-clustering method, we use two meta-clustering algorithms: either k-means based or latent class based. The k-means based meta-clustering either uses CCEA convergent k-means clustering or k-means highest reproducibility replicate (Sawtooth software CCEA, version 3). The latent class based meta-clustering uses Latent Gold's latent class clustering. See Figure 5 for an overview of the industry study design.

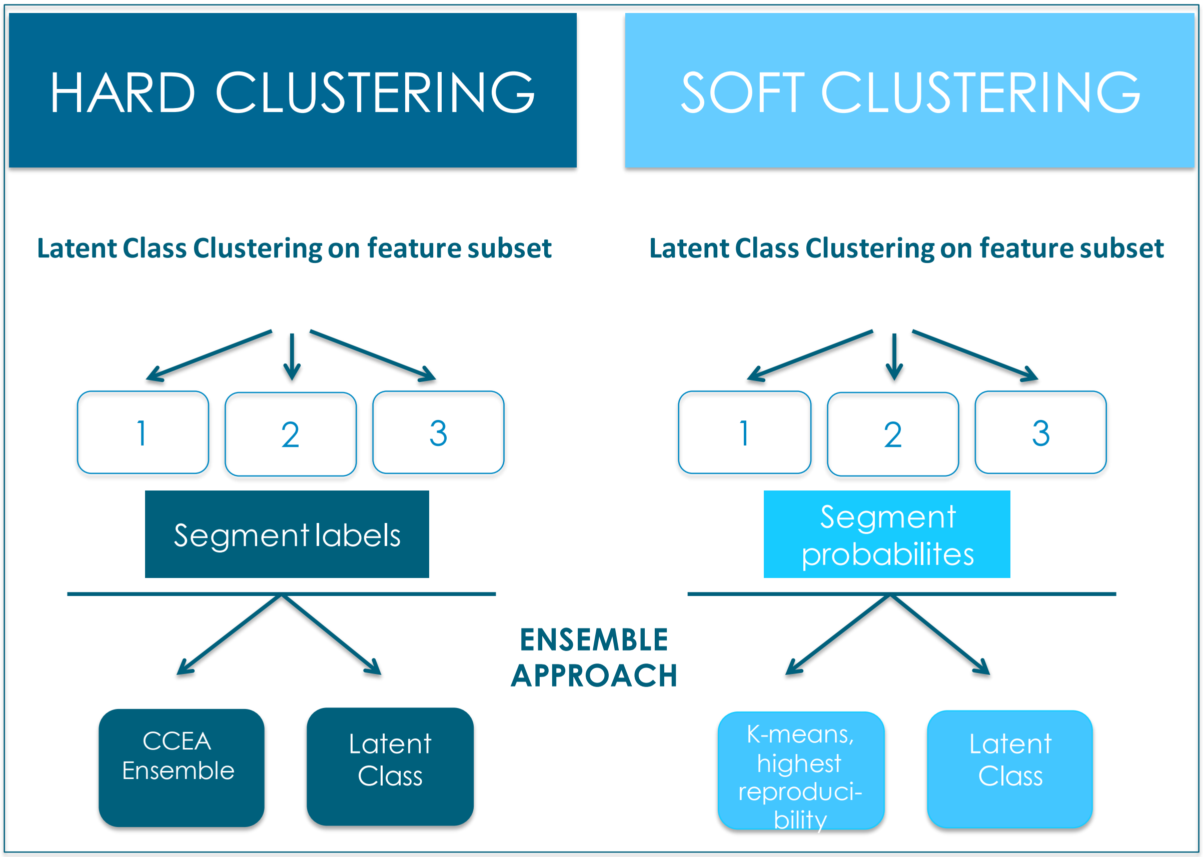


Figure 5: Industry experiment's study design

The two "hard" cluster ensembles are compared with two "soft" cluster ensembles using the previously discussed evaluation criteria: stability, integrity (homogeneity and heterogeneity), accuracy and size. The stability of the segmentations on a particular industry data set is calculated as the average of the similarity indices for a 3 to 4 cluster solution, a 4 to 5 cluster solution and a 5 to 6 cluster solution. To determine the accuracy, we either used a multinomial logit algorithm or a discriminant analysis to predict the segments using the survey questions.

Figure 6 shows the evaluation for stability, integrity (homogeneity and heterogeneity) and accuracy. In each of the subfigures, the bar *color* and bar *line style* represent a different industry data set. The bar *size* represents the difference in number of wins between soft vs hard cluster ensembles for a particular industry data set.

The number of wins definition for stability and accuracy is different than for homogeneity and heterogeneity. A method wins in terms of stability if it has a higher average similarity index. A method wins in terms of accuracy if it has a higher Adjusted Rand Index on the best cluster solution as determined by the client from 3 to 6 cluster solutions. We calculate the total number of wins for soft cluster ensembles as the number of times k-means soft and LC soft win from the hard cluster ensembles; CCEA hard and Latent Class hard. Likewise, the number of wins for hard cluster ensembles is calculated as the number of times CCEA hard and LC hard win from the soft cluster ensembles; k-means soft and Latent Class soft.

i.e. k-means soft versus (CCEA hard and LC hard)
LC soft versus (CCEA hard and LC hard) } # wins soft

CCEA hard versus (k-means soft and LC soft)
LC hard versus (k-means soft and LC soft) } # wins

The definition of the number of wins for homogeneity and heterogeneity is defined differently than for stability. Firstly, the homogeneity and heterogeneity are determined for each of the KPIs. Subsequently, the number of wins per method are recorded as the number of KPIs for which the method has the best performance (lowest scatter/lowest total separation) after which we sum the total number of wins for k-means soft and Latent Class soft as the number of wins for soft cluster ensembles. Analogously, we sum the total number of wins for CCEA hard and Latent Class hard as the number of wins for hard cluster ensembles.

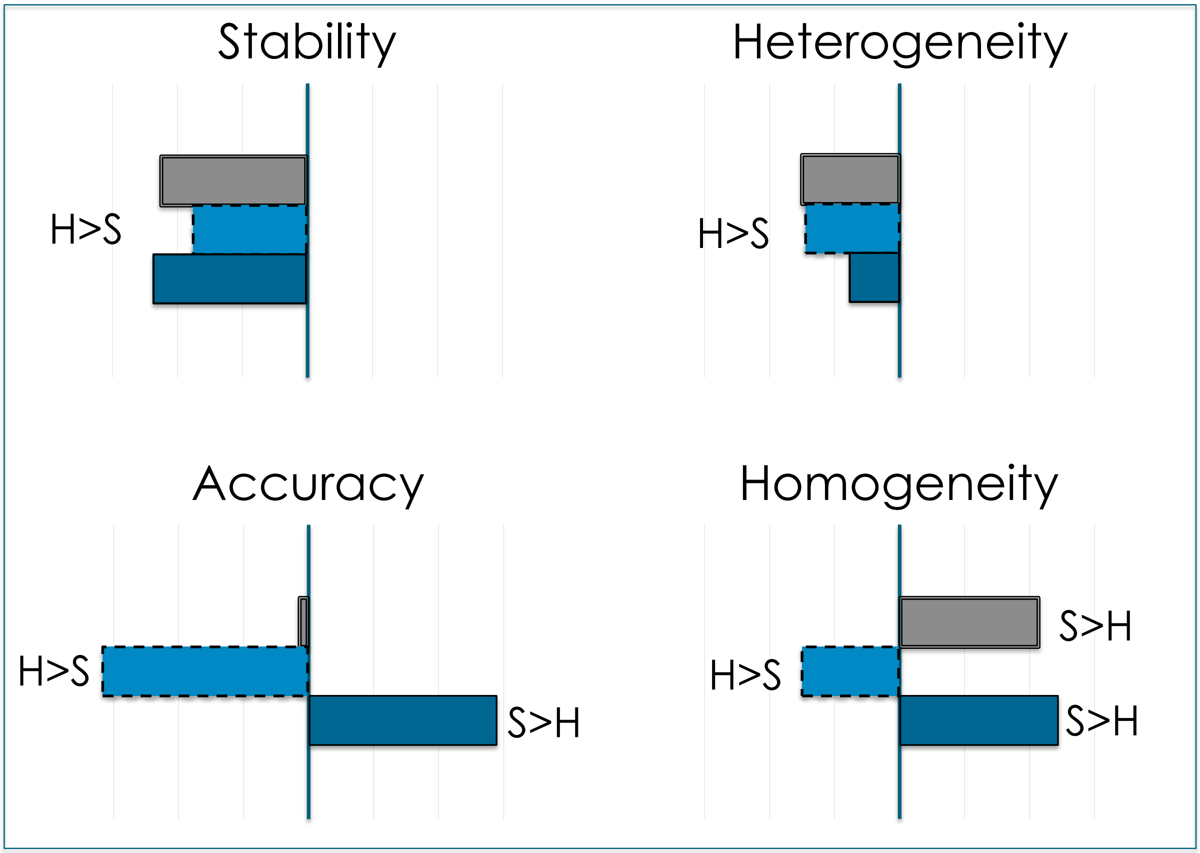


Figure 6: Stability, integrity (homogeneity and heterogeneity) and accuracy on three industry data sets (color bars). Bar length represents difference in #wins for soft cluster ensembles S vs #wins for hard cluster ensembles H.

On all three industry data sets the hard cluster ensemble provides more stable segmentations than the soft cluster ensembles and results in more distinct segments (#wins Hard > #wins Soft). However, we get mixed evidence on whether a soft or hard cluster ensemble is preferred in terms of how homogeneous and predictable the segments are.

The two soft and two hard cluster ensembles are also compared with respect to the size of the segments. The average uniformity deviation over the three industry data sets is substantially larger for the latent class-based ensembles than for the k-means based ensembles: 8.28 percentage points vs 4.9 percentage points. The three graphs in Figure 7 show for each industry data set the % of persons in the segments ordered by decreasing segment size. The latent-class based ensembles have a higher tendency to build small segments than the k-means based ensembles. For that reason, latent-class based ensembles are no longer considered in subsequent comparisons between soft and hard cluster ensembles.

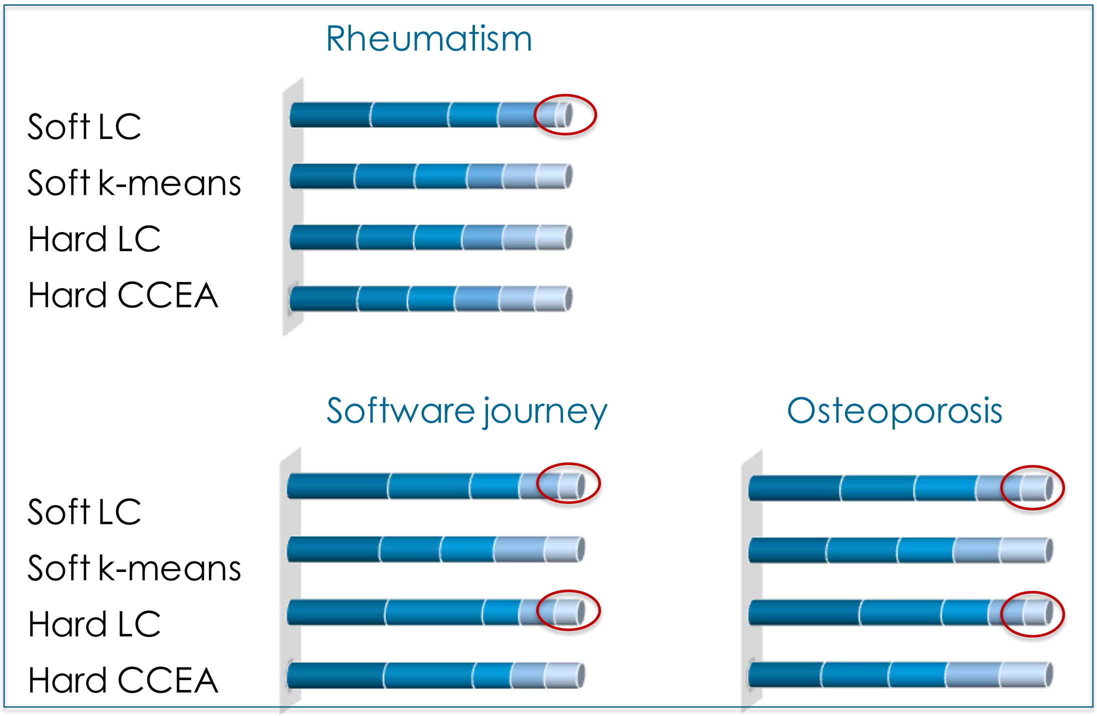


Figure 7: Percentage persons in the segments ordered by decreasing segment size for the three industry data sets. Soft ensembles have a higher tendency to build smaller segments (see red circles).

Simulation Study

The industry experiments do not provide clear guidance on when to use soft versus hard cluster ensembles. Therefore a simulation study is set up to indicate when to use a soft or hard cluster ensemble given that you know:

1) the similarity of the clustering solutions combined by the ensemble
2) and the confidence in the cluster allocation of the clustering solution.

The simulation design is a 2-by-2 full-factorial design as shown in Figure 8.

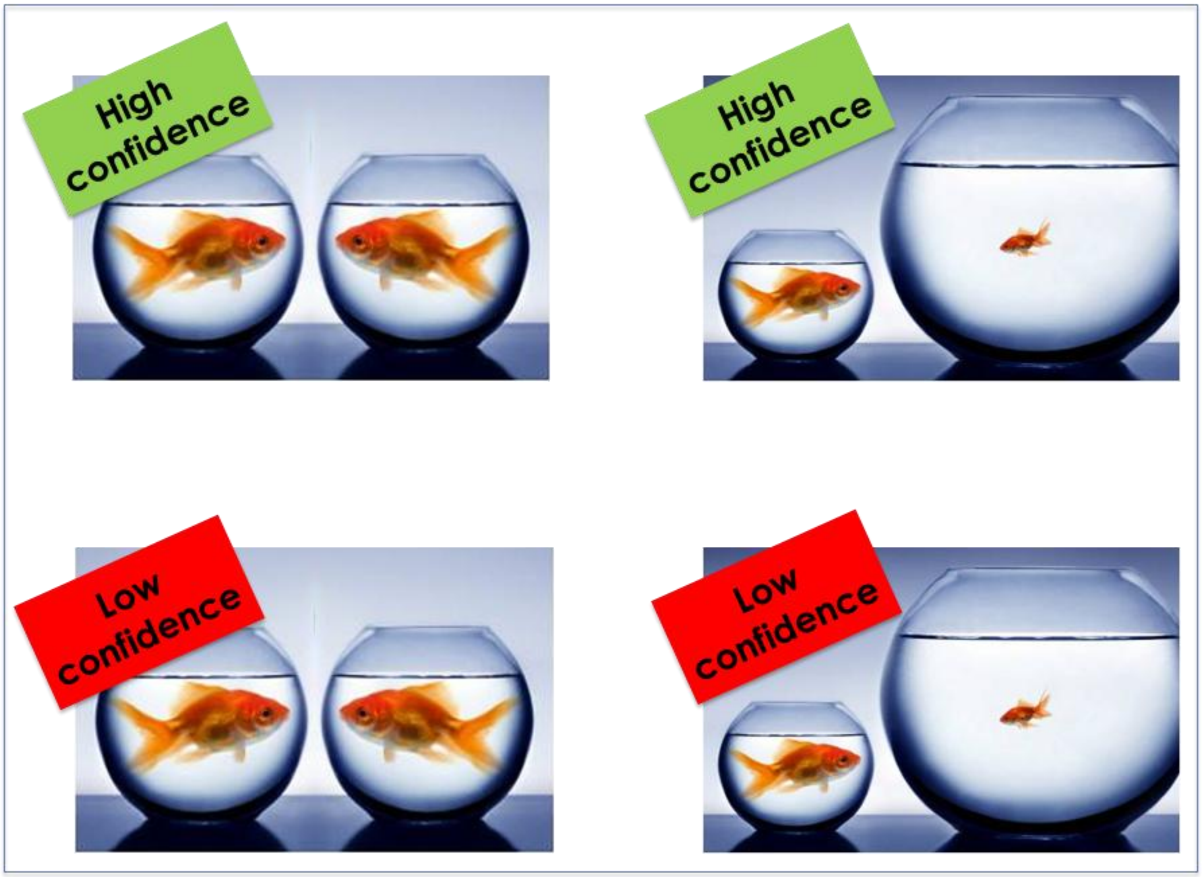


Figure 8: Simulation 2-by-2 design

Simulation data generation

The simulated data is generated according to a simulation design with fixed and variable factors.

The fixed factors of the simulation design are the number of clusterings combined, the number of segments in each clustering and the segment proportions. Each cluster ensemble combines 10 clusterings with each clustering having 400 observations belonging to one of four segments. All 10 clusterings are class balanced cluster solutions with 100 persons per segment. We choose balanced clusterings in the simulated data as typically, when selecting the best clustering for a survey theme, you would choose the most balanced clustering solution too.

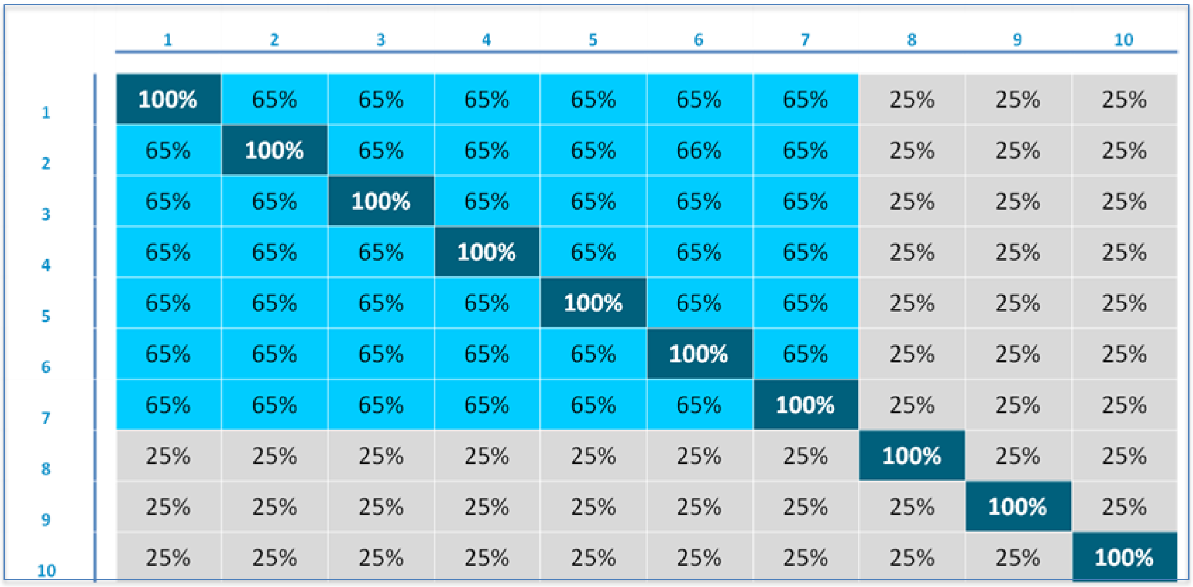
We first create single clustering solutions by generating cluster membership probabilities. Later on, we use the segment labels or probabilities of 10 of these single clustering solutions as input to respectively a hard or soft cluster ensemble.

Each clustering solution generated either has high confidence or low confidence in a person's segment membership. In the 'high' confidence clustering solution, a person belongs for 95% probability to his allocated segment. In the 'low' confidence clustering solution, a person belongs for 70% to his allocated segment. To generate high or low confidence clustering solutions we generate probabilities to belong to each of the four segments using a Dirichlet distribution. To create a 'high' confidence clustering solution, we generate 100 observations for each of the following Dirichlet distributions: Dirichlet(95,2,2,2), Dirichlet (2,95,2,2), Dirichlet (2,2,95,2) and Dirichlet (2,2,2,95). To create a 'low' confidence clustering solution, we generate 100 observations for each of the following Dirichlet distributions: Dirichlet(70,10,10,10), Dirichlet (10,70,10,10), Dirichlet (10,10,70,10) and Dirichlet (10,10,10,70). We use the Markov Chain Monte Carlo package in R to draw the samples (Martin, Quinn, and Park, 2003-2012).

Besides high or low confidence, the other variable factor is the similarity of the clusterings combined by the soft or hard cluster ensemble. The clustering solutions combined by the ensemble could be strongly similar or weakly similar.

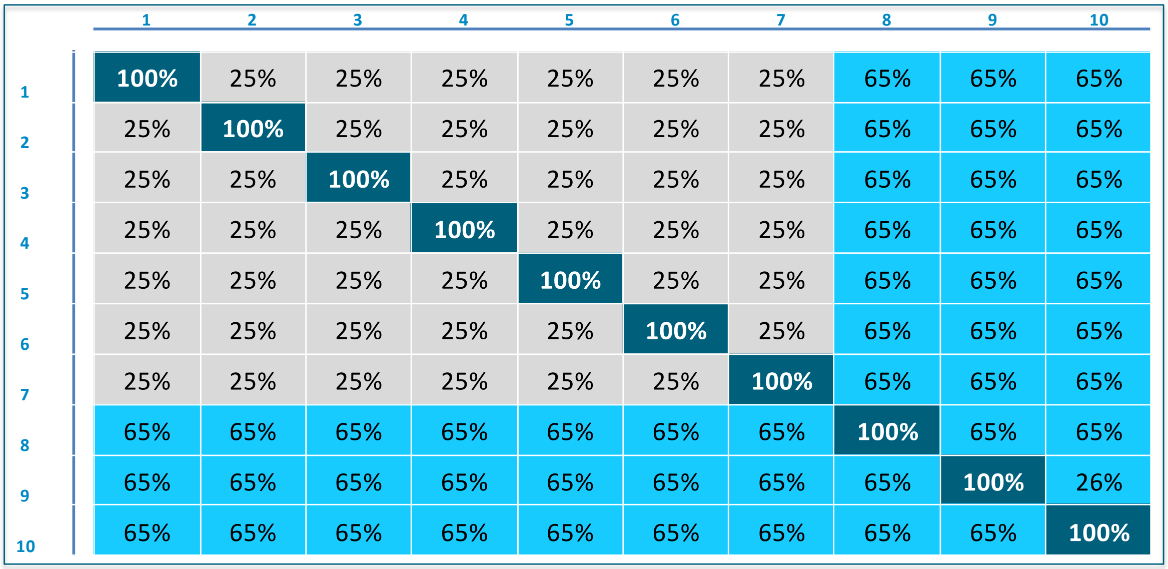
The *strong similarity* simulated data set combines 10 solutions (either all high confidence or all low confidence) of which 7 have a similarity index of approximately 0.65. Originally, the single clustering solutions drawn from the appropriate Dirichlet distribution have a similarity index of 1. That is, a person belonging to a given segment in solution X has a 100% probability to belong to that same segment in solution Y. By randomly permuting the probabilities for 25% of the observations per segment across 7 original clustering solutions, the similarity index between these 7 solutions is reduced to 0.65. Likewise, for 3 out of 10 solutions with a perfect match all observations were randomly permuted resulting in a similarity index of approx. 25%. The median similarity index over 10 solutions is 65%. Figure 9 shows the similarity matrix between the 10 clustering solutions for a strong similarity simulated data set.

The *weak similarity* simulated data set combines 10 solutions (either all high confidence or all low confidence) of which 7 have a similarity index of approximately 25%. Therefore, all observations for 7 out of 10 solutions with a perfect match across the 7 solutions were randomly permuted. For 3 out of 10 solutions with a perfect match 25% of the observations per segment were randomly permuted resulting in a similarity index of approximately 65%. The median similarity index over 10 solutions is 25%. See Figure 9 for the similarity matrix. Figure 10 shows the similarity matrix between the 10 clustering solutions for a weak similarity simulated data set.



	1	2	3	4	5	6	7	8	9	10
1	100%	65%	65%	65%	65%	65%	65%	25%	25%	25%
2	65%	100%	65%	65%	65%	66%	65%	25%	25%	25%
3	65%	65%	100%	65%	65%	65%	65%	25%	25%	25%
4	65%	65%	65%	100%	65%	65%	65%	25%	25%	25%
5	65%	65%	65%	65%	100%	65%	65%	25%	25%	25%
6	65%	65%	65%	65%	65%	100%	65%	25%	25%	25%
7	65%	65%	65%	65%	65%	65%	100%	25%	25%	25%
8	25%	25%	25%	25%	25%	25%	25%	100%	25%	25%
9	25%	25%	25%	25%	25%	25%	25%	25%	100%	25%
10	25%	25%	25%	25%	25%	25%	25%	25%	25%	100%

Figure 9: similarity between 10 strongly similar clustering solutions



	1	2	3	4	5	6	7	8	9	10
1	100%	25%	25%	25%	25%	25%	25%	65%	65%	65%
2	25%	100%	25%	25%	25%	25%	25%	65%	65%	65%
3	25%	25%	100%	25%	25%	25%	25%	65%	65%	65%
4	25%	25%	25%	100%	25%	25%	25%	65%	65%	65%
5	25%	25%	25%	25%	100%	25%	25%	65%	65%	65%
6	25%	25%	25%	25%	25%	100%	25%	65%	65%	65%
7	25%	25%	25%	25%	25%	25%	100%	65%	65%	65%
8	65%	65%	65%	65%	65%	65%	65%	100%	65%	65%
9	65%	65%	65%	65%	65%	65%	65%	65%	100%	26%
10	65%	65%	65%	65%	65%	65%	65%	65%	65%	100%

Figure 10: similarity between 10 weakly similar clustering solutions

For each of the four different data settings of the 2-by-2 design we generate 10 data sets (i.e. 10 data sets for the high confidence-strong similarity, 10 data sets for the high confidence-low similarity, 10 data sets for the low confidence-high similarity and 10 data sets for the low confidence-low similarity scenario). Consequently, we can make 10 comparisons between the soft cluster ensemble and the hard cluster ensemble, rather than just one.

Both hard and soft cluster ensembles are estimated using Sawtooth Software's CCEA with minimum 3 and maximum 6 groups (segments). The hard cluster ensembles use 'Ensemble, consensus solution' as clustering method and start from the simulation data segment labels. The soft cluster ensembles use 'k-means highest reproducibility replicate' clustering method and start from the simulation data segment membership probabilities.

Simulation Results

The soft and hard cluster ensembles are evaluated using the previously discussed evaluation criteria: stability, integrity (homogeneity and heterogeneity) and size. The soft and hard cluster ensembles cannot be evaluated with respect to accuracy because the simulation data only contains segment probabilities and membership data but no predictors of persons' segment membership. The evaluation metrics are averaged over the 10 simulation data sets created per simulation scenario (high/low confidence * high/low similarity).

Soft ensembles are more stable when combining highly similar clusterings

The stability of the segmentations on a particular simulation data set is calculated as the average of the similarity indices for a 3 to 4 cluster solution, a 4 to 5 cluster solution and a 5 to 6 cluster solution. This average similarity index is calculated for the 10 simulation data sets in a given design setting (e.g. high confidence, strong similarity) and averaged.

Figure 11 shows the evaluation of soft and hard cluster ensembles on the simulation data with respect to stability. The vertical axis portrays the difference between the average similarity index for a soft and a hard cluster ensemble. A soft ensemble is more stable when it combines strongly similar clustering solutions. The effect of the clustering confidence on the stability of the segmentation can be neglected.

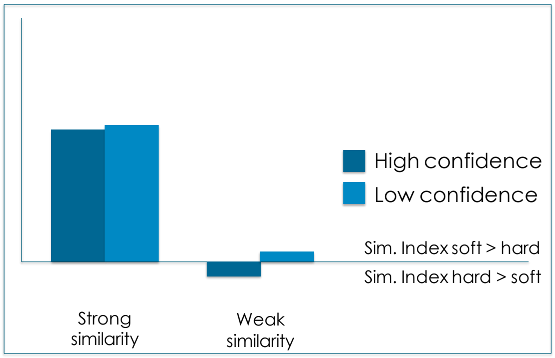


Figure 11: Soft ensembles are more stable

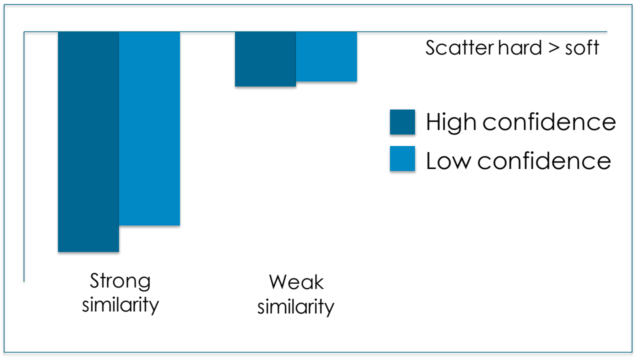


Figure 12: Soft ensembles provide more homogeneous segments

Soft ensembles provide more homogeneous segments

The soft and hard cluster ensembles are compared with respect to how homogeneous the produced segments are. Segments are homogenous when the persons within a segment are highly similar. The lower the scatter the more homogeneous the segments are. Although the hard and soft cluster ensembles provide three to six cluster solutions, the scatter is calculated for a four cluster solution only as all the clusterings combined by the ensemble are also four cluster solutions.

To define on which dimensions the segments in the simulation data sets should be homogenous is not as straightforward as it was for the industry data sets. On the industry data sets we defined the segment homogeneity with respect to KPIs. On the simulation data sets however we cannot measure how similar the observations are on KPIs or other descriptive variables. After all, these simulation data sets only contain segment membership probabilities and the segment label. Therefore, we define the homogeneity in terms of the similarity of the segment membership probabilities. As the ensembles combine 10 clusterings with varying degree of similarity, it does not make sense to calculate the homogeneity on the segment membership probabilities of all 10 clustering solutions. We should calculate the homogeneity in terms of the segment probabilities of the clusterings that are strongly similar. Recall that the *strongly similar* simulated data sets combine 7 clusterings that have a high similarity index of approximately 0.65. Also recall that the *weak similarity* simulated data sets combine 10 solutions (either all high confidence or all low confidence) of which only 3 have a high similarity index of approximately 25%. In short, for the strongly similar simulated data sets we measure the homogeneity of the segment observations on the membership probabilities of the 7 highly similar clusterings. For the weakly similar simulated data sets the homogeneity of the segment observations is calculated on the membership probabilities of the three highly similar clusterings.

Figure 12 shows the evaluation of the soft and hard cluster ensembles with respect to homogeneity. Recall, that the scatter has been averaged over the 10 simulated datasets in a given design setting (e.g. high confidence, strong similarity). The vertical axis portrays the difference between the average scatter index for a soft and a hard cluster ensemble. The lower the scatter the more homogeneous the segments are. Hard cluster ensembles have more scatter so produce less homogeneous segments. A soft ensemble's segments are more homogeneous especially when it combines highly similar clusterings. When the ensemble combines clusterings that are weakly similar, it does not really matter whether segment labels or segment probabilities are used.

Hard cluster ensembles provide more distinct segments

Hard and soft cluster ensembles are compared with respect to segment heterogeneity. Segments are heterogeneous when persons from different segments are highly dissimilar. The lower the total separation metric, the more heterogeneous the segments are. Although the hard and soft cluster ensembles provide three to six cluster solutions, the total separation is calculated for a four cluster solution only as all the clusterings combined by the ensemble are also four cluster solutions.

Analogous to the calculation of the homogeneity on the simulation data sets, the segment heterogeneity is defined using the segment probabilities of the highly similar clusterings combined by the ensemble. That is, for the strongly similar simulated data sets the total separation is calculated on the membership probabilities of the 7 highly similar clusterings. For the weakly similar simulated data sets the heterogeneity of the segment observations is calculated on the membership probabilities of the three highly similar clusterings.

Figure 13 shows the evaluation of the soft and hard cluster ensembles with respect to heterogeneity. Recall, that the total separation has been averaged over the 10 simulated datasets in a given design setting (e.g. high confidence, strong similarity). The vertical axis portrays the difference between the average total separation for a soft and a hard cluster ensemble. A lower total separation the more heterogeneous the segments are. Hard cluster ensembles have lower total separation so produce more distinct segments.

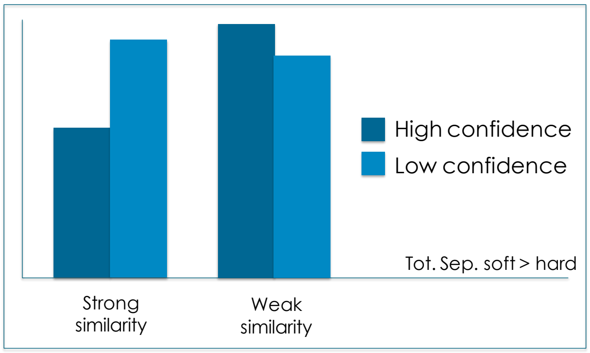


Figure 13: Hard cluster ensembles provide more more distinct segments

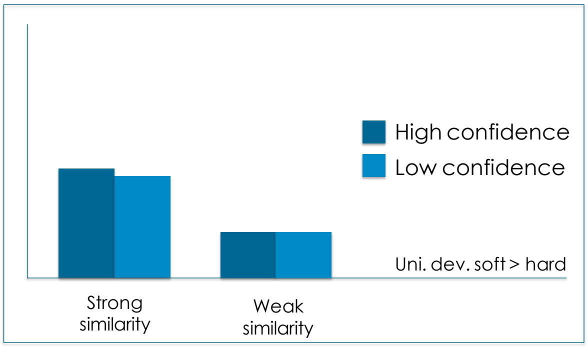


Figure 14: Hard cluster ensembles provide balanced segments

Hard cluster ensembles provide more balanced segments

The soft and hard cluster ensembles are also compared with respect to the size of the segments. The average uniformity deviation over the 10 simulation data sets in a given scenario is calculated for a four segment solution. Hence, we compare how much the segment proportions deviate from 25% (absolute deviation). The soft ensembles show larger uniformity deviations (Figure 14) so provide segments that tend to be unequal in size. In conclusion, hard cluster ensembles provide more balanced segments.

Hard Ensembles Are Recommended for Business Applications

Cluster ensembles improve industry segmentations by answering both the “who” and “why”. They allow managers to know why customers in a segment behave as they do and who these segment members are. This paper assessed the business value of “hard” versus “soft” cluster ensembles. “Hard” cluster ensembles combine segment labels from multiple cluster solutions into a meta-segmentation. “Soft” cluster ensembles adopt a more granular approach by using the detailed cluster membership probabilities as input to the meta-segmentation.

The industry experiments provided mixed evidence. The simulation study however reveals that hard cluster ensembles provide more distinct segments and better balanced segments than soft cluster ensembles do. However, soft cluster ensembles excel in the stability and homogeneity of the segments. Having done many segmentation projects and talked to many users of segmentations, we believe that finding segments which are (1) very distinct and therefore often easier to target and (2) not too big nor too small is more important than having homogenous clusters. Therefore, we recommend the hard cluster approach.

The simulation results surprised us in multiple ways.

Firstly, we expected that using the rich segment membership data would deliver better segmentations than using the coarser segment labels. Often the more rich the data input, the better the results. This does not seem to hold true. Using the cluster probabilities might lead to information overload in the sense that the “signal-to-noise” ratio is bad. When highly similar clusterings are combined, the segment membership data acts as noise. The latter is apparent from lower segmentation heterogeneity for soft cluster ensembles than for hard cluster ensembles.

Secondly, we expected that using the cluster membership probabilities would be beneficial to the segmentation as it reveals the clustering’s confidence in the segment allocation. We therefore expected large differences between the low and high confidence simulation results. The results show that combining low or high confidence clusterings does not really influence the quality of the segmentation except for segment heterogeneity.

Several avenues are open for further research. Only two levels of clustering similarity were used in this paper: high (0.65(7)0.25(3)) and low (0.25(7)0.65(3)). As the simulation results portray that the clustering similarity is the main factor influencing the difference between soft and hard cluster ensembles, future research should investigate more clustering similarity levels. Furthermore, in this paper the confidence level is fixed within a clustering (a particular clustering assigns all respondents with the *same* probability to their segment) and between the ensemble clusterings (all clusterings assign respondents to their segment with the same level of confidence). For example, in the low confidence scenario all 10 clusterings assign all respondents with 70% to their segment. Future research could have a clustering confidence level that is variable within and/or between clusterings. Finally, it might be worth looking into hybrid cluster ensembles using hard segment labels for some respondents, yet soft segment membership probabilities for other respondents.

REFERENCES

Fred, A.L.N. and Jain, A.K. (2002), Data clustering using evidence accumulation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 835-850.

Halkidi, M.,Vazirgiannis M. and Batistakis, Y. (2000), Quality Scheme Assessment in the Clustering Process, *Proc. of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, 265-276.

Hubert, L. and Arabie, P. (1985), Comparing partitions, *Journal of Classification*, 193-218.

Lange T., Roth, V., Braun L. and Buhmann J.M. (2004), Stability-based validation of Clustering solutions, *Neural Computation*, 16, 1299-1323.

Martin, A., Quinn, K.M. And Park, J.H. (2003-2012), Markov Chain Monte Carlo Package (MCMCpack), R software.

Nieweglowski, L., CLV package (2007), R software.

Strehl, A. and Gosh, J. (2002), Cluster ensembles – A knowledge reuse framework for combining partitionings, *Journal of Machine Learning Research*, 3, 583-618.

Topchy, A., Jain A.K., and Punch W. (2004), A mixture model for clustering ensembles, In: *Proceedings of the SIAM International Conference on Data Mining*, Michigan State University, USA.

HOW LOW CAN YOU GO?: TOWARD A BETTER UNDERSTANDING OF THE NUMBER OF CHOICE TASKS REQUIRED FOR RELIABLE INPUT TO MARKET SEGMENTATION

JANE TANG
ANDREW GRENVILLE
VISION CRITICAL

ABSTRACT

Tang and Grenville (2010) examined the tradeoff between the number of choice tasks and the number of respondents for CBC in the era of on-line panels. The results showed that respondents become less engaged in later tasks. Therefore, increasing the number of choice tasks brings limited improvement in the model's ability to predict respondents' behavior, and actually decreases model sensitivity and consistency.

We recommended that CBC practitioners who utilize on-line panel samples should take pains to keep their respondents happy and engaged. This can be accomplished by minimizing the length and complexity of the choice tasks based on the model requirements. Whenever possible, we also recommend increasing the sample size (lower sampling error) to compensate for the lower precision in modeling that results from a smaller number of tasks.

However, increasing sample size to compensate for fewer tasks does not help when a high level of precision at the individual level is required. This happens most often when the results from CBC are to be used as the basis for a market segmentation.

In this paper we take a closer look at this problem, using seven industry studies involving both MaxDiff and CBC experiments. We develop segments using convergent cluster analysis (CCA) from the reduced choice experiment and compare the results to the segments developed using the full choice information.

We find that the clustering process itself is often unstable even with a large number of choice tasks providing full information. We conclude that using a slightly smaller number of tasks is not harmful to the segmentation process. In fact, under most conditions, a choice experiment using only 10 tasks is sufficient for segmentation purposes.

We would like to thank our colleague Ray Poynter for his comments and contributions.

1. INTRODUCTION

Johnson & Orme (1996) was an early paper that addressed the suitable length of a CBC experiment. Their analysis was based on 21 data sets contributed by Sawtooth Software CBC users. Since this was before the age of Internet panels, the data were collected through CAPI or via mail-in disks using pre-recruited samples. The authors determined that respondents can answer at least 20 choice tasks without degradation in data quality, though their analysis was limited to pooled estimation (aggregate logit).

Hoogerbrugge & van der Wagt (2006) was a later paper that addressed this issue. It focused on holdout board hit rate prediction. They found that 10-15 tasks are generally sufficient for the majority of studies. The increase in hit rates beyond that number is minimal.

Markowitz & Cohen (2001) dealt with the tradeoff of larger sample vs. longer conjoint task in a Hierarchical Bayes (HB) framework. They approached this problem using simulated data of various sample sizes and model complexity. They concluded that the predictive value of the HB model is not greatly improved by increasing the sample size; more choice tasks per respondent are better than more respondents.

As a result, it is common for CBC practitioners to employ about 15 tasks in their conjoint experiment. Going beyond 15 tasks brings minimal improvement to the precision of the HB models. Among practitioners, the focus is often on pushing the respondents hard with long choice tasks while keeping the sample size small.

Today, the widespread use of internet panels throws some doubt on these results. The simulated respondents used by Markowitz & Cohen never got bored or annoyed, and the quality of their responses did not decline as the number of tasks increased. Real people aren't so patient. In the verbatim feedback from our panelists we see repeated complaints about the length and repetitiveness of choice tasks.

Longer conjoint exercises result in longer questionnaires; the effects of which have been demonstrated (Galesic and Bosnjak (2009), Rathod & la Bruna (2005) and Cape (2010)) to have numerous negative effects, including:

- Lower response rates
- Lower completion rates
- Greater rates of cheating
- Lower data quality

Rathod & la Bruna advise us that "If researchers work to keep surveys shorter, it will not only help ensure response quality, but it will also make for more motivated and responsive respondents."

Tang and Grenville (2010) examined the tradeoff between choice tasks and sample size, particularly with respect to on-line panels. Our findings demonstrated that respondents become less engaged in later tasks. Increasing the number of choice tasks brings limited improvement in the model's ability to predict respondents' behavior, and actually decreases model sensitivity and consistency.

We recommend that CBC practitioners who use on-line panels should take pains to keep their panelists happy and engaged. This can be accomplished by minimizing the length and complexity of choice tasks based on the model requirements. Whenever possible, consider increasing the sample size (lower sampling error) to compensate for the lower precision in modeling resulting from the smaller number of tasks.

However, increasing sample size to compensate for fewer tasks does not help when a high level of precision at the individual level is required. This happens most often when the results from CBC are used as the basis for a market segmentation.

2. Our Hypothesis

We hypothesize that while lower precision may negatively affect the segmentation process, the end result may not in fact be any worse. While a long choice task section allows for more precise individual-level estimates in terms of predictability, the deterioration in data quality during the later tasks may lead one to question the quality of the estimates in terms of sensitivity and consistency.

More importantly, for the purpose of segmentation, the actual realized gain in heterogeneity from the later choice tasks may not be as significant as previously thought; the segments derived from the clustering process using the reduced choice information (i.e. removing the later choice tasks) may perform as well as those using the full choice information.

3. Data sets

We collected seven industry data sets. We would like to thank Bryan Orme from Sawtooth Software and Kees van der Wagt from SKIM for their help in locating these data sets.

To facilitate the analysis, we required that each data set must have a large sample (n=800 or more) and a fairly large number of tasks (10 or more) in the CBC/MaxDiff section.

Name	Category	Type of Task	N= Sample Size	T= # of Tasks	a= # of alter Natives	# of factors	# of Parameters estimated
VC1	CPG	MaxDiff	914	14	5	1	23
VC2	Food	MaxDiff	992	14	5	1	23
VC3	Financial Services	CBC, Dual Response	2,465	10	3	9	25
SKIM1	--	CBC, Dual Response	978	16	4	8	31
SKIM2	--	CBC, Dual Response	4,051	15	5	4	27
SKIM3	--	CBC, Dual Response	1,139	15	3	7	24
ST1	--	CBC, Dual Response	1,202	12	5	3	20

4. The process

To assess the validity of segments derived from the reduced choice information, we first established segments using the full choice information. This was accomplished by running an HB model first and then creating the segments through a convergent cluster analysis (CCA) procedure.

We recognize there are many different ways to establish market segments based on choice information. We feel the process we are using here (HB first, following by clustering) is truest to the issue at hand, which is the potential loss of heterogeneity within the HB process resulting from reduced choice tasks.

For each of the seven data sets we created an HB model using all **T** choice tasks. The resulting respondent level beta utilities were then put through an anti-logit transformation.

$$f(\beta) = \frac{e^{\beta}}{1+e^{\beta}} .$$

It is well known that the raw beta utility score from the HB process is tied to a scale parameter; the better the model fit, the larger the magnitude of the estimate. The transformed betas do not suffer from this problem; they are much easier to work with, having a range of between 0 and 1, with an expected value of 0.5.

The transformed betas were put through a CCA run. The best reproducing cluster solutions (using the best starting points, for 2 through 8 clusters) were retained for further study. These are our target clusters.

All the betas were used in clustering (rather than using only the few betas most relevant to the research objectives) since the research objectives are unknown. This also takes the "art" factor out of clustering and allows us to compare the more "generic" clusters across multiple data sets.

The process was then repeated, using **T**-1 choice tasks instead (eliminating the last choice task). The resulting clusters were considered to be secondary clusters.

We compared the secondary clusters to the target clusters using the following criteria,

- Reproducibility: How well can the secondary clusters be matched to the target clusters?

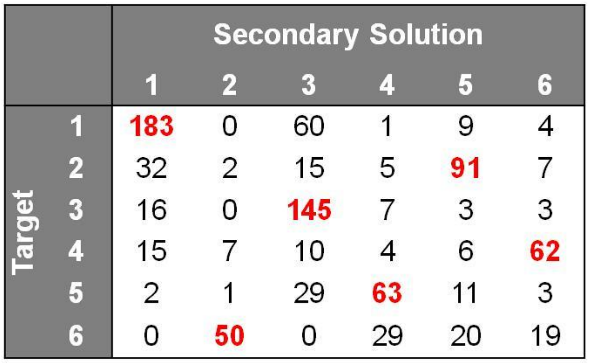
- MAE: How different are the secondary cluster profiles compared to their corresponding target cluster profiles?

We then repeated the same calculation using **T**-2 choice tasks, **T**-3 choice tasks, etc …, removing the last choice task from the HB estimation each time.

4.1. Calculating Reproducibility

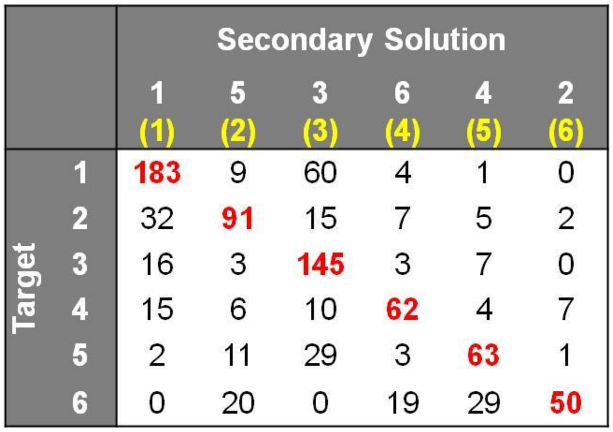
To calculate reproducibility we followed the procedure used to match two sets of cluster solutions documented in the section 6 of the Sawtooth Software CCEA manual. However, in this case we were comparing two sets of cluster solutions calculated using different data sets from the same respondents.

For example, the following shows how the secondary 6-cluster solution is cross tabulated against the target 6-cluster solution.



		Secondary Solution					
		1	2	3	4	5	6
Target	1	**183**	0	60	1	9	4
	2	32	2	15	5	**91**	7
	3	16	0	**145**	7	3	3
	4	15	7	10	4	6	**62**
	5	2	1	29	**63**	11	3
	6	0	**50**	0	29	20	19

We can then exchange pairs of columns to derive the following table where the sum of the diagonal elements is maximized.



		Secondary Solution					
		1	5	3	6	4	2
		(1)	(2)	(3)	(4)	(5)	(6)
Target	1	**183**	9	60	4	1	0
	2	32	**91**	15	7	5	2
	3	16	3	**145**	3	7	0
	4	15	6	10	**62**	4	7
	5	2	11	29	3	**63**	1
	6	0	20	0	19	29	**50**

The adjusted reproducibility can be calculated as follows:

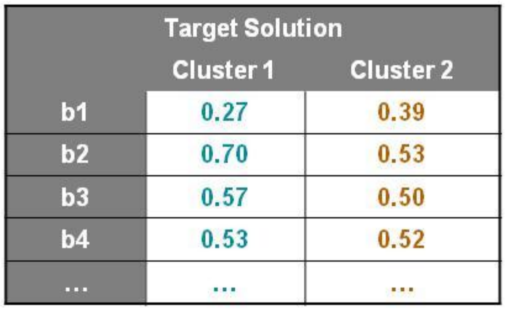
$$a = \frac{k \times r - 1}{k - 1}$$

r = the unadjusted reproducibility proportion, which is the sum of the diagonal elements over the total sample size.
k = the number of clusters
a = the adjusted reproducibility value

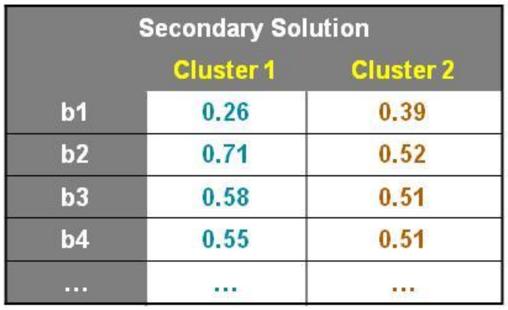
In this example, we calculated an adjusted reproducibility of 58%.

4.2. Calculating MAE

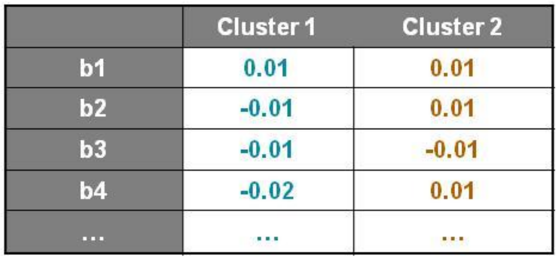
The permutated columns in the above example also allow us to match each cluster within the secondary solution to its closest cousin in the target solution – these are the numbers in brackets near the top of the table. We can then compute the differences between cluster centers (using the transformed betas) for each cousin pair.



	Target Solution	
	Cluster 1	Cluster 2
b1	0.27	0.39
b2	0.70	0.53
b3	0.57	0.50
b4	0.53	0.52
...	...	...



	Secondary Solution	
	Cluster 1	Cluster 2
b1	0.26	0.39
b2	0.71	0.52
b3	0.58	0.51
b4	0.55	0.51
...	...	...



	Cluster 1	Cluster 2
b1	0.01	0.01
b2	-0.01	0.01
b3	-0.01	-0.01
b4	-0.02	0.01
...	...	...

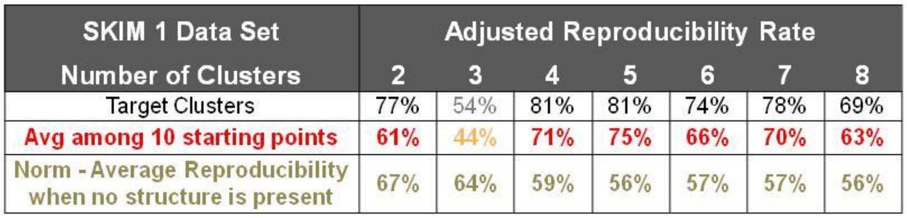
The Mean Absolute Error (MAE) in the example above is 0.01. MAE measures the similarity between the corresponding cluster centers between the target and secondary solution.

5. Pass/Fail Tests

All the data sets were put through two pass/fail tests, one for reproducibility and one for MAE.

5.1. Test for reproducibility

The clustering process itself is inherently unstable. While you can always find clusters, not all the solutions are worthwhile. Using the reproducibility norms established in section 6 of the Sawtooth Software CCEA manual, we assessed the reproducibility of the secondary clusters only when a target cluster was considered to be stable. This happens when the adjusted reproducibility (across the various 10 starting points) from the target cluster is higher than the norm average when no structure is present.



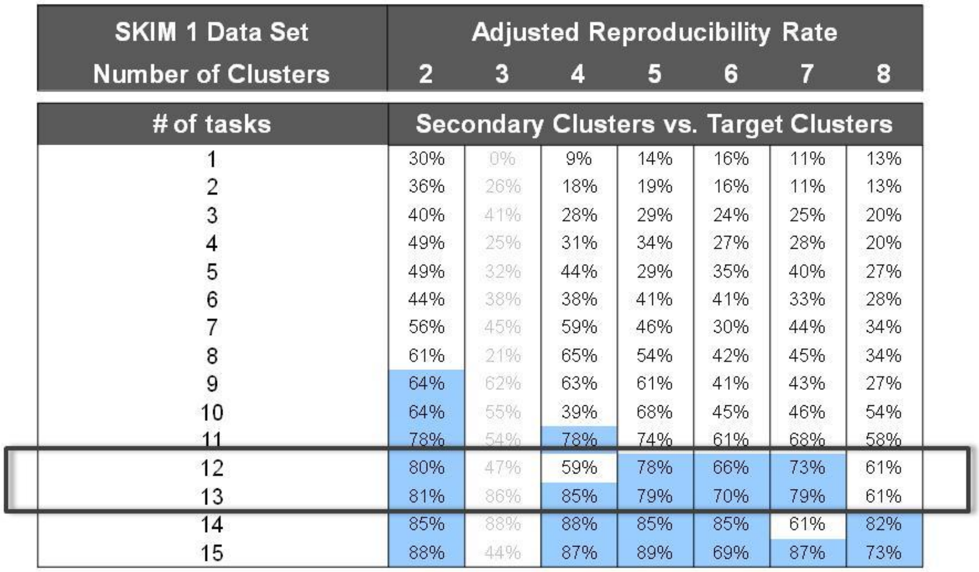
SKIM 1 Data Set Number of Clusters	Adjusted Reproducibility Rate 2	3	4	5	6	7	8
Target Clusters	77%	54%	81%	81%	74%	78%	69%
Avg among 10 starting points	61%	44%	71%	75%	66%	70%	63%
Norm - Average Reproducibility when no structure is present	67%	64%	59%	56%	57%	57%	56%

In the example above, the target 3-cluster solution does not meet our criterion. All secondary 3-cluster solutions from this data set are excluded from our consideration for the test of reproducibility and MAE.

For the target clusters, the adjusted reproducibility rates for all 10 starting points are retained. We use the average as a measure of how well we can expect the clusters to reproduce using the full choice information.

A secondary cluster solution is judged to have passed the test of reproducibility if,

- It has a good adjusted reproducibility, i.e. 20% higher than the norm.
  OR
- It reproduces as well as one would expect from the other starting points using data from the full HB run.



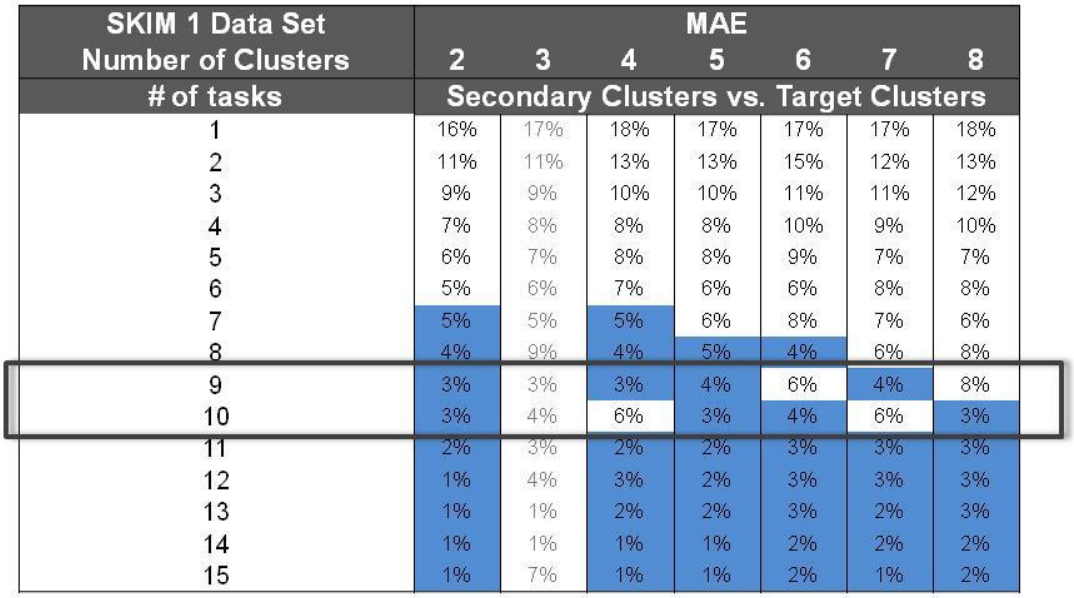
SKIM 1 Data Set Number of Clusters	Adjusted Reproducibility Rate 2	3	4	5	6	7	8
# of tasks	Secondary Clusters vs. Target Clusters						
1	30%	0%	9%	14%	16%	11%	13%
2	36%	26%	18%	19%	16%	11%	13%
3	40%	41%	28%	29%	24%	25%	20%
4	49%	25%	31%	34%	27%	28%	20%
5	49%	32%	44%	29%	35%	40%	27%
6	44%	38%	38%	41%	41%	33%	28%
7	56%	45%	59%	46%	30%	44%	34%
8	61%	21%	65%	54%	42%	45%	34%
9	64%	62%	63%	61%	41%	43%	27%
10	64%	55%	39%	68%	45%	46%	54%
11	78%	54%	78%	74%	61%	68%	58%
12	80%	47%	59%	78%	66%	73%	61%
13	81%	86%	85%	79%	70%	79%	61%
14	85%	88%	88%	85%	85%	61%	82%
15	88%	44%	87%	89%	69%	87%	73%

In general, the more tasks that are included in the HB process, the better we can reproduce the target clusters. The larger cluster solutions (e.g. 7 or 8) are harder to reproduce than the smaller cluster solutions (i.e. 2 or 3). In the example here, you need about 12-13 tasks to reproduce the target cluster (total number of tasks T=16).

However, there are exceptions, in some cases, the reproducibility rate drops as more tasks are used.

5.2. Test for MAE

The transformed (anti-logit of) beta has an expected value of 50%. So a 5% MAE allows for 10% tolerance for the test. If a secondary solution has less than 5% MAE, it is considered to have passed the MAE test.



SKIM 1 Data Set Number of Clusters	MAE 2	3	4	5	6	7	8
# of tasks	Secondary Clusters vs. Target Clusters						
1	16%	17%	18%	17%	17%	17%	18%
2	11%	11%	13%	13%	15%	12%	13%
3	9%	9%	10%	10%	11%	11%	12%
4	7%	8%	8%	8%	10%	9%	10%
5	6%	7%	8%	8%	9%	7%	7%
6	5%	6%	7%	6%	6%	8%	8%
7	5%	5%	5%	6%	8%	7%	6%
8	4%	9%	4%	5%	4%	6%	8%
9	3%	3%	3%	4%	6%	4%	8%
10	3%	4%	6%	3%	4%	6%	3%
11	2%	3%	2%	2%	3%	3%	3%
12	1%	4%	3%	2%	3%	3%	3%
13	1%	1%	2%	2%	3%	2%	3%
14	1%	1%	1%	1%	2%	2%	2%
15	1%	7%	1%	1%	2%	1%	2%

Similar to what we observed in the test of reproducibility, the MAE decreases as more tasks are used in the HB process. MAE's are bigger as the number of clusters increases.

Again, there are exceptions – sometimes having more tasks in the HB process leads to larger MAE.

It seems to be easier (i.e. fewer tasks required) to have secondary solutions with profiles similar to the target clusters. Only about 9-10 tasks are required in the example here. We see this pattern repeated in other data sets as well. The MAE test is less stringent than the reproducibility test.

6. Summary of Tests

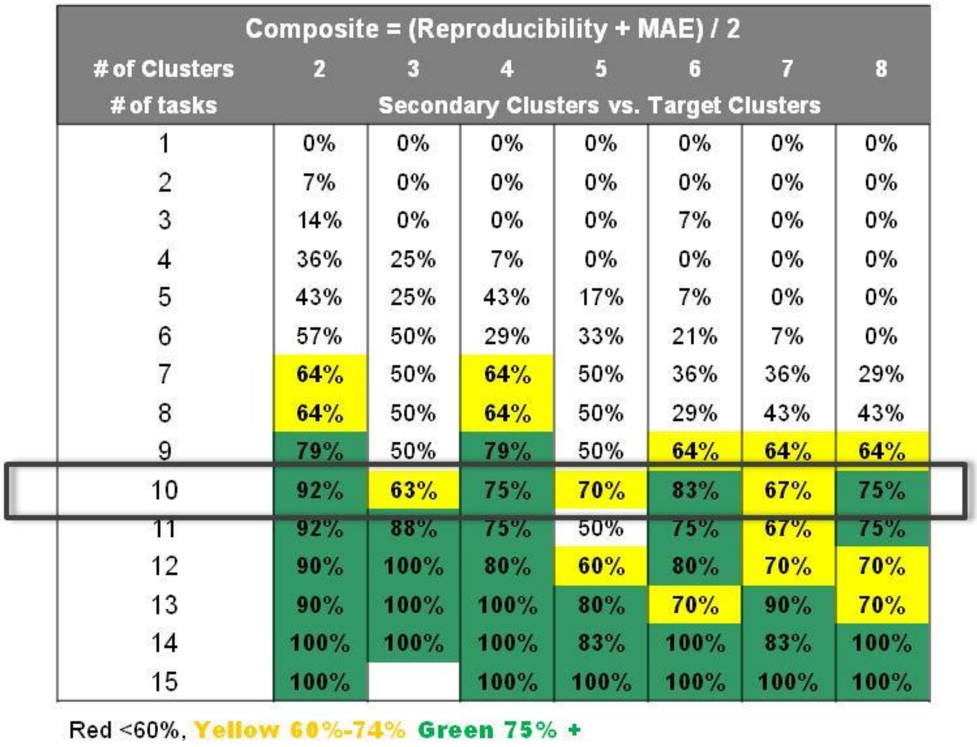
We count the number of passes across 7 data sets for both tests, and compute the composite score as follows:

$$Composite\ Score = \frac{Number\ of\ \ Passes\ In\ \mathrm{Re}\ producibility\ Test + Number\ of\ \ Passes\ In\ MAE\ Test}{Total\ Number\ of\ \ Tests}$$

Note here that we averaged the number of occasions the secondary solutions passed the tests we devised. We are not averaging the actual reproducibility or MAE statistics, rather the number of occasions the numbers are favorable enough to have passed the test.

Only those target clusters considered stable are used in the calculation.

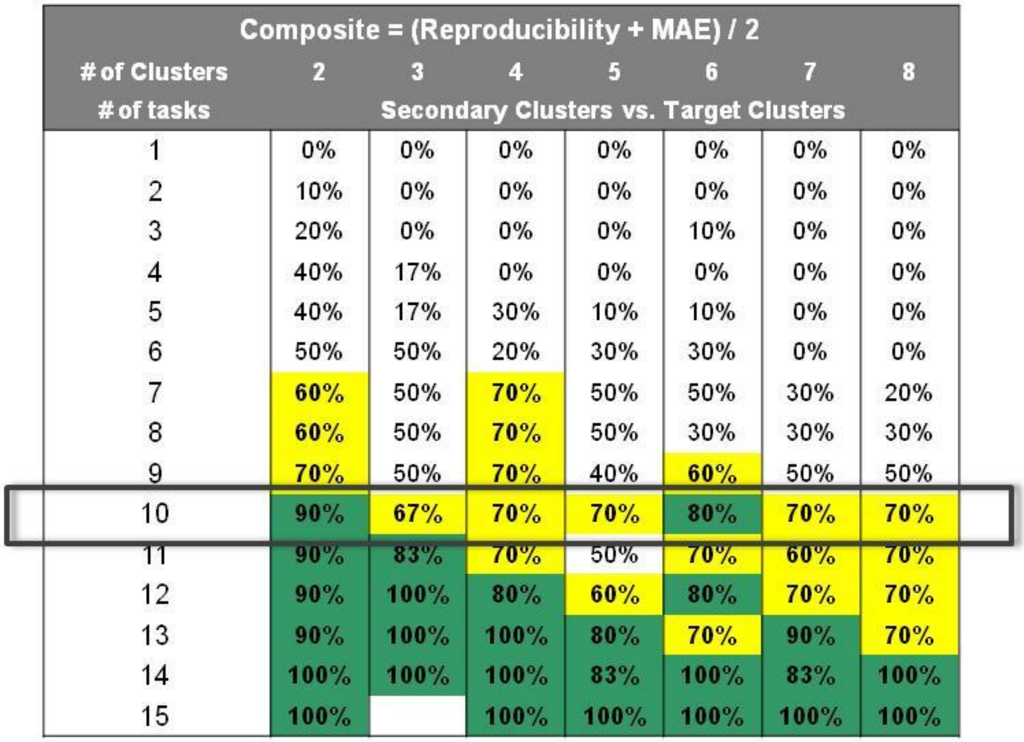
It appears that only 10 choice tasks are needed to successfully replicate the target solution.



Composite = (Reproducibility + MAE) / 2							
# of Clusters	2	3	4	5	6	7	8
# of tasks	Secondary Clusters vs. Target Clusters						
1	0%	0%	0%	0%	0%	0%	0%
2	7%	0%	0%	0%	0%	0%	0%
3	14%	0%	0%	0%	7%	0%	0%
4	36%	25%	7%	0%	0%	0%	0%
5	43%	25%	43%	17%	7%	0%	0%
6	57%	50%	29%	33%	21%	7%	0%
7	64%	50%	64%	50%	36%	36%	29%
8	64%	50%	64%	50%	29%	43%	43%
9	79%	50%	79%	50%	64%	64%	64%
10	92%	63%	75%	70%	83%	67%	75%
11	92%	88%	75%	50%	75%	67%	75%
12	90%	100%	80%	60%	80%	70%	70%
13	90%	100%	100%	80%	70%	90%	70%
14	100%	100%	100%	83%	100%	83%	100%
15	100%		100%	100%	100%	100%	100%

Red <60%, Yellow 60%-74% Green 75% +

Since two of the data sets have 10 and 12 tasks, we are concerned that this might bias the results. Excluding these two data sets, however, does not change our conclusion.



Composite = (Reproducibility + MAE) / 2							
# of Clusters	2	3	4	5	6	7	8
# of tasks	Secondary Clusters vs. Target Clusters						
1	0%	0%	0%	0%	0%	0%	0%
2	10%	0%	0%	0%	0%	0%	0%
3	20%	0%	0%	0%	10%	0%	0%
4	40%	17%	0%	0%	0%	0%	0%
5	40%	17%	30%	10%	10%	0%	0%
6	50%	50%	20%	30%	30%	0%	0%
7	60%	50%	70%	50%	50%	30%	20%
8	60%	50%	70%	50%	30%	30%	30%
9	70%	50%	70%	40%	60%	50%	50%
10	90%	67%	70%	70%	80%	70%	70%
11	90%	83%	70%	50%	70%	60%	70%
12	90%	100%	80%	60%	80%	70%	70%
13	90%	100%	100%	80%	70%	90%	70%
14	100%	100%	100%	83%	100%	83%	100%
15	100%		100%	100%	100%	100%	100%

7. Conclusion

Our findings show that while decreasing the number of choice tasks (to about 10) leads to slightly less precise HB estimates, the heterogeneity retained in the reduced task is sufficient for follow up analysis, such as segmentation, where precision at the individual level is required.

The clustering process is often inherently unstable even with large number of choice tasks. Using more tasks in the HB process leads to clusters that more closely resemble the target solutions, i.e. they reproduce more accurately, with smaller MAE. Segmentation with a larger number of clusters requires more choice tasks.

Under most conditions, a conjoint exercise involving ten tasks appears to be sufficient for segmentation purposes.

But why ten tasks? We have two conjectures.

One: within the online survey environment, the first 10 (or so) tasks gave you sufficient information for segmentation. Fewer tasks are not sufficient, more tasks add little.

Two: we suspect that there is a loss of engagement in later tasks. While additional questions (beyond 10) might theoretically help, the deterioration in quality of responses negates their value.

There are many different ways to derived segments from choice data. While we used HB followed by clustering, we are not saying that is always the best way of accomplishing that objective. We used this process because it is best suited for our purpose; demonstrating the impact of the loss of precision and heterogeneity when it comes to segmentation.

We recommend that CBC practitioners who utilize on-line panel samples should take pains to keep their panelists happy and engaged. This can be accomplished by minimizing the length and complexity of the choice tasks. Whenever possible, we recommend increasing sample size (lower sampling error) to compensate for the lower precision in modeling resulting from the smaller number of tasks.

In market segmentation studies, or any studies that require very precise individual level estimates, we should consider limiting the number of choice tasks to no more than ten; striking a balance between model precision and respondent fatigue.

REFERENCES

Cape P. (2010), “Questionnaire Length, Fatigue Effects & Response Quality: Revisited” ARF re:think 2010.

CCEA User Manual v3 (2008) Sawtooth Software.

Galesic M. and Bosniak M. (2009), “Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey” Public Opinion Quarterly Volume 73, Issue 2 pp. 349-360.

Hoogerbrugge, M. and van der Wagt, K. (2006) “How Many Choice Tasks Should We Ask?” Sawtooth Software Conference Proceedings.

Johnson, R. and Orme, B. (1996), “How Many Questions Should You Ask In Choice-Based Conjoint Studies?” ART Forum Proceedings.

Markowitz, P. and Cohen, S. (2001), “Practical Consideration When Using Hierarchical Bayes Techniques?” ART Forum Proceedings.

Rathod S and LaBruna A (2005), “Questionnaire length and fatigue effects - does size really matter?” ESOMAR, Conference on Panel Research, Budapest.

Tang, J. and Grenville, A. (2010), “How Many Questions Should You Ask in CBC Studies? – Revisited Again” Sawtooth Software Conference Proceedings.

"THE INDIVIDUAL CHOICE TASK THRESHOLD" NEED FOR VARIABLE NUMBER OF CHOICE TASKS

PETER KURZ
TNS INFRATEST FORSCHUNG GMBH
STEFAN BINNER
BMS - MARKETING RESEARCH + STRATEGY

SUMMARY

This paper reflects on the fact that CBC is the most widely used conjoint methodology in our industry, and that the authors' experience shows that fewer tasks are often all that is needed (fewer than typical norms in the industry). Furthermore, with the widespread use of online panel sample, panel providers and respondents are pushing for fewer tasks in CBC questionnaires. By observing individual respondents completing CBC surveys, we show that many respondents reach a point in the CBC survey where they disengage and begin to give data of questionable quality—a point we call the "Individual Choice Task Threshold (ICT)." We re-analyzed twelve commercial CBC data sets to see if quantitative measures could be used to somehow detect the threshold point at which respondents disengaged. The idea is that if somehow this could be detected in real-time, then the survey could not ask any more tasks, respondent burden could be reduced, and fewer overall choice tasks would need to be asked for the sample, while maintaining equal or better results. Even if this were never possible, we might benefit from finding the ICT after the fact and at least dropping the later tasks from the analysis.

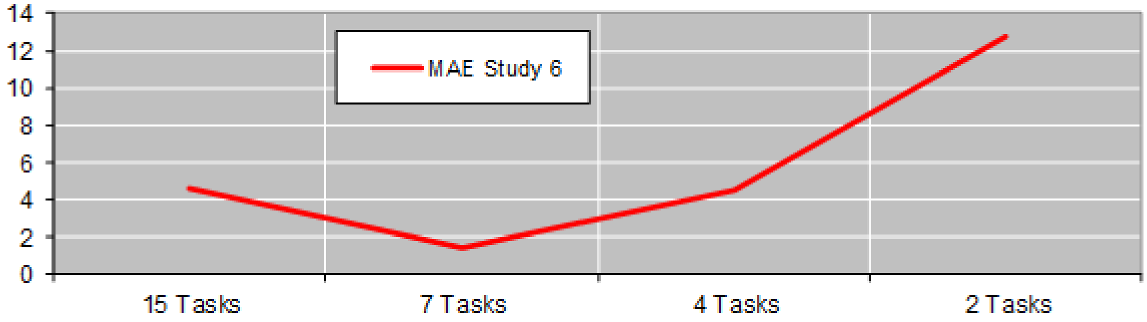
The first part of the analysis focuses on measures of internal fit to the tasks used in estimation (RLH), which is used also in an indexed form to take the decrease of the RLH measure into account when more information (higher number of choice tasks) is available. Using the existing data sets, it could be demonstrated that strategically throwing away 38% of the choice tasks would not lead to a very large decrease in the predictive quality of the CBC models. The results show clearly that most respondents use simplification strategies in later compared to earlier tasks. Based on these findings, the last part of the paper presents some ideas as to how future development could take the ICT in online surveys into account to shorten the interview. But, up to now there are no solution available to do this in real-time and apply it during data collection. Further research and more computational power are needed to solve this problem.

Introduction

Since its introduction, CBC (choice based conjoint) has become the most established method for conjoint analysis. Thousands of conjoint studies are conducted by the market research industry every year. Especially with the application of HB, alternative specific designs, disproportional sampling and long experience, CBC usually leads to high-quality results, successful projects and satisfied clients.

Despite the success researchers have with CBC, there are some observations which may be troubling:

- **Are fewer tasks enough?**
  Prior research (Kurz, Binner; 2010) showed some surprising results: Not using the last choice tasks did not automatically decrease the quality of simulations! As following example shows, the MAE[8] was unchanged even when we reduced the number of tasks from 15 to only four.



- **Do we bedevil respondents?**
  Answering conjoint choice tasks is a complex and monotonous exercise. More and more often one can hear or read comments from respondents complaining about conjoint surveys. Many find the interviews too long and complex. Especially in view of the consolidation of online panels, one more often finds experienced respondents who express their dissatisfaction. As a consequence, higher incentives are necessary to keep them answering.

- **Is there a choice task threshold?**
  Observations during personal conjoint interviews regularly show that some respondents mentally opt-out very early and some apply simplification in the choice tasks, while others remain engaged until the end of the exercise. This indicates that every respondent has her own "Choice Task Threshold" after which she is no longer contributing useful answers.

- **How can they be so fast?**
  As is widely known (Johnson, Orme; 1996) the answer times for the last choice tasks tend to decrease, which could be an indication of simplification or less concentration.

[8] MAE (Mean Absolute Error) calculations in this paper are done at the individual-level. For the holdout, the selected alternative gets 100% share, and the other alternatives 0%. For predictions, the share of preference (logit) rule is used. The average of the absolute differences between predicted and holdout shares is computed as MAE for each respondent, and the results averaged across respondents.

HYPOTHESES FOR THIS PAPER

Based on the above observations we formulated the following hypotheses:

H1: The observed task simplification of some respondents leads to less accurate prediction of market shares.

H2: If respondents lose their interest in the conjoint exercise, this at least results in more noisy data.

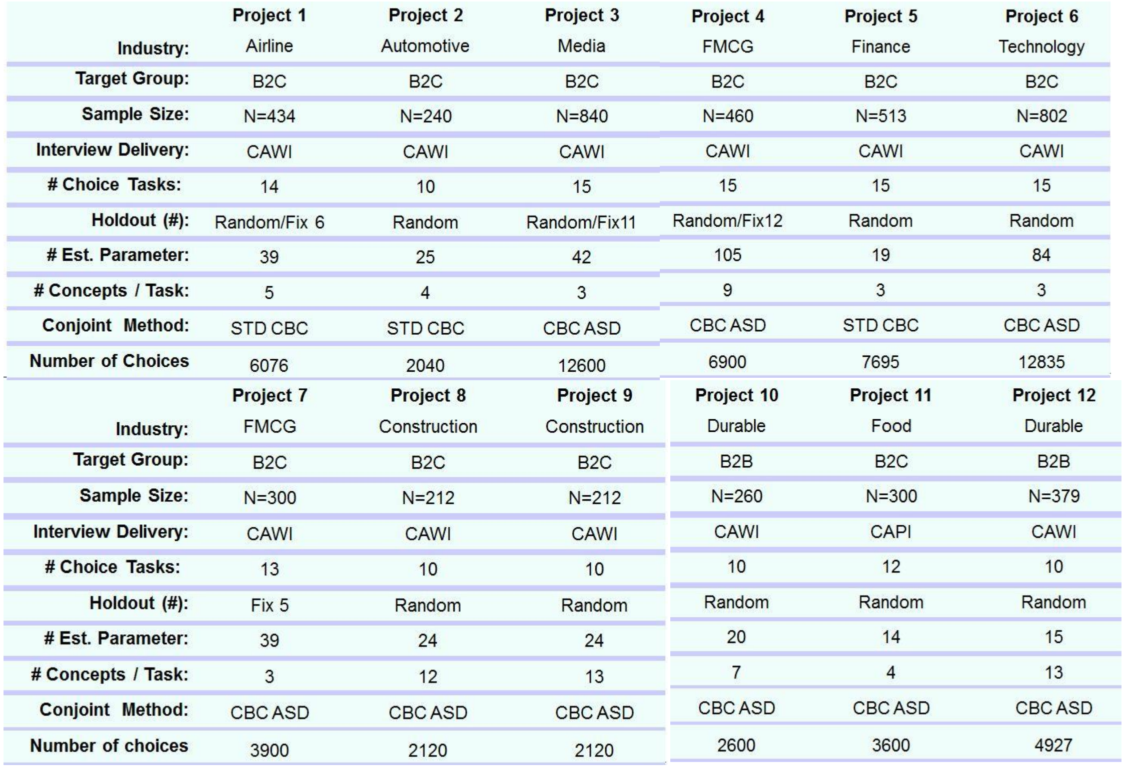
H3: Our results can therefore be improved when we use only those "good" tasks in which respondents were concentrating and paid attention.

H4: The shift in price sensitivity of later tasks (Johnson, Orme; 1996) could also be avoided by using "good" tasks only.

H5: Finally, interview time could be shortened, if we are able to avoid "bad tasks" during the interview, leading to reduced cost or a larger sample for the same cost.

BASIS FOR THE ANALYSIS

For the purposes of this paper 12 commercial studies with a total of 4,952 respondents and 67,413 choice tasks covering nearly all industries and topics were analyzed. The 12 studies included "brand + price" CBC designs as well as designs with many attributes (between 10-15 choice tasks and 14 to 105 estimated parameters). The selected studies cover all main market fields such as industrials, durables, and FMCG, in both B2B and B2C markets. They were conducted in all parts of the world using state-of the-art computer aided interview delivery and mainly recruited from online panels. Finally all 12 studies were developed in Sawtooth Software's CBC software, with all analysis done with HB using default settings.



	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6
Industry:	Airline	Automotive	Media	FMCG	Finance	Technology
Target Group:	B2C	B2C	B2C	B2C	B2C	B2C
Sample Size:	N=434	N=240	N=840	N=460	N=513	N=802
Interview Delivery:	CAWI	CAWI	CAWI	CAWI	CAWI	CAWI
# Choice Tasks:	14	10	15	15	15	15
Holdout (#):	Random/Fix 6	Random	Random/Fix11	Random/Fix12	Random	Random
# Est. Parameter:	39	25	42	105	19	84
# Concepts / Task:	5	4	3	9	3	3
Conjoint Method:	STD CBC	STD CBC	CBC ASD	CBC ASD	STD CBC	CBC ASD
Number of Choices	6076	2040	12600	6900	7695	12835

	Project 7	Project 8	Project 9	Project 10	Project 11	Project 12
Industry:	FMCG	Construction	Construction	Durable	Food	Durable
Target Group:	B2C	B2C	B2C	B2B	B2C	B2B
Sample Size:	N=300	N=212	N=212	N=260	N=300	N=379
Interview Delivery:	CAWI	CAWI	CAWI	CAWI	CAPI	CAWI
# Choice Tasks:	13	10	10	10	12	10
Holdout (#):	Fix 5	Random	Random	Random	Random	Random
# Est. Parameter:	39	24	24	20	14	15
# Concepts / Task:	3	12	13	7	4	13
Conjoint Method:	CBC ASD	CBC ASD	CBC ASD	CBC ASD	CBC ASD	CBC ASD
Number of choices	3900	2120	2120	2600	3600	4927

Figure 1: Overview of studies analyzed

All 12 studies were analyzed, task by task, in terms of

- Individual time needed to answer the tasks
- Individual RLH reached per task (absolute and indexed)
- Individual utilities derived with different numbers of tasks
- Hit Rates and MAE for fixed and random tasks with different numbers of tasks (for consistency, second-to-last random task was used for fit measurement).

The computations were monitored by four standard measures, which are based on the calculation of the probability of each respondent choosing as she did on each task, by applying a logit model using current estimates of each respondent's part worths. The likelihood is the product of those probabilities over all respondents and tasks. Usually the logarithm of this measure is used, the “log likelihood”. Measures were:

Percent Certainty - indicates how much better the solution is than chance, as compared to a “perfect" solution.
Root Likelihood (RLH) - nth root of the likelihood, where n is the total number of choices made by all respondents in all tasks. RLH is therefore the geometric mean of the predicted probabilities.
Variance - the average of the current estimate of the variances of part worths, across respondents.
RMS - the root mean square of all part worth estimates, across all part worths and over all respondents.

Answer Times and RLH

As expected, the average answer times decline to a low and stable level within a few choice tasks:

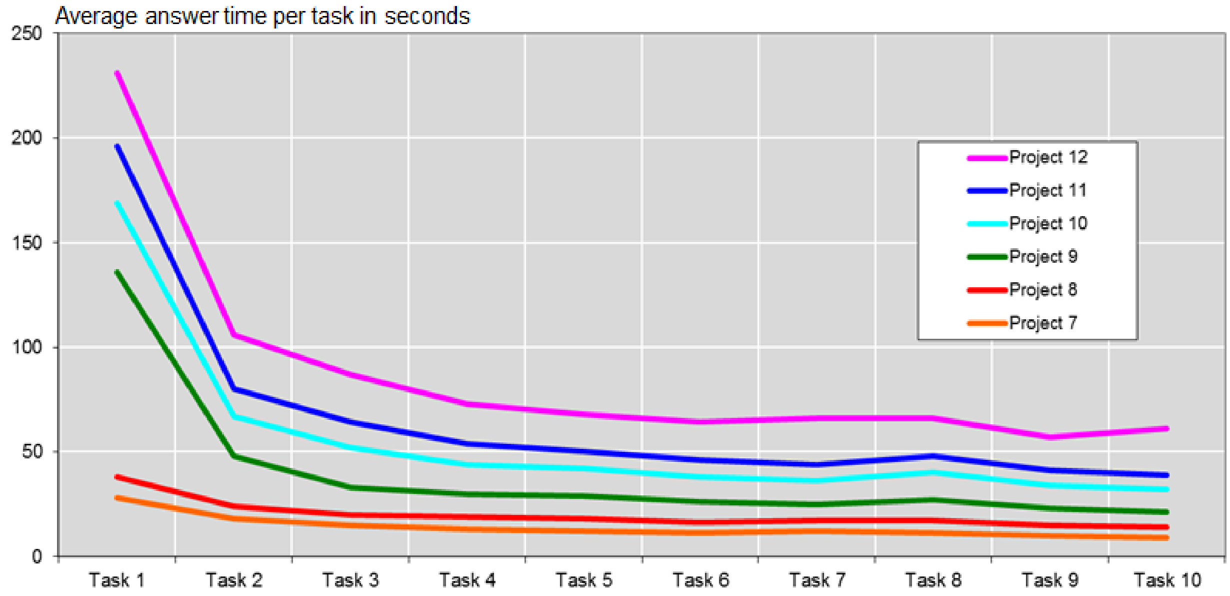


Figure 2: Average answer times throughout conjoint questionnaires

However there was no significant correlation between the average answer time and the RLH level. In addition, changes in individual answering time between choice tasks had no significant influence on the RLH of the answer. Slower or faster answer times for individual choice tasks are not good indicators for less attention, shortcutting or reaching the choice task threshold, as the times can be influenced by respondents' environment or the utility balance of the tasks. Furthermore the aggregated RLH was stable or even declining from choice task to choice task:

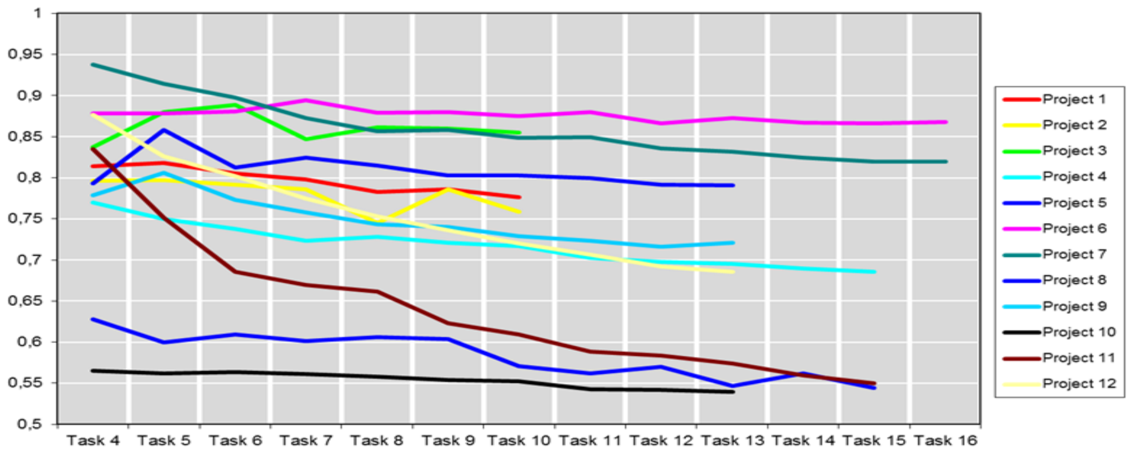


Figure 3: Aggregated RLH throughout conjoint questionnaires

This result was confirmed by all four measures for monitoring the HB process, as shown in the next four graphs. In these figures, and in most later results, the data plotted for "Task 4" is the overall figure from an analysis using only the first 4 tasks for each respondent, while that for

"Task 5" is from a re-analysis using only the first 5 tasks for each respondent, etc. In effect, "Task n" reflects a possible level of the ICT (individual choice threshold) that is being evaluated. We did not analyze using three or fewer tasks.

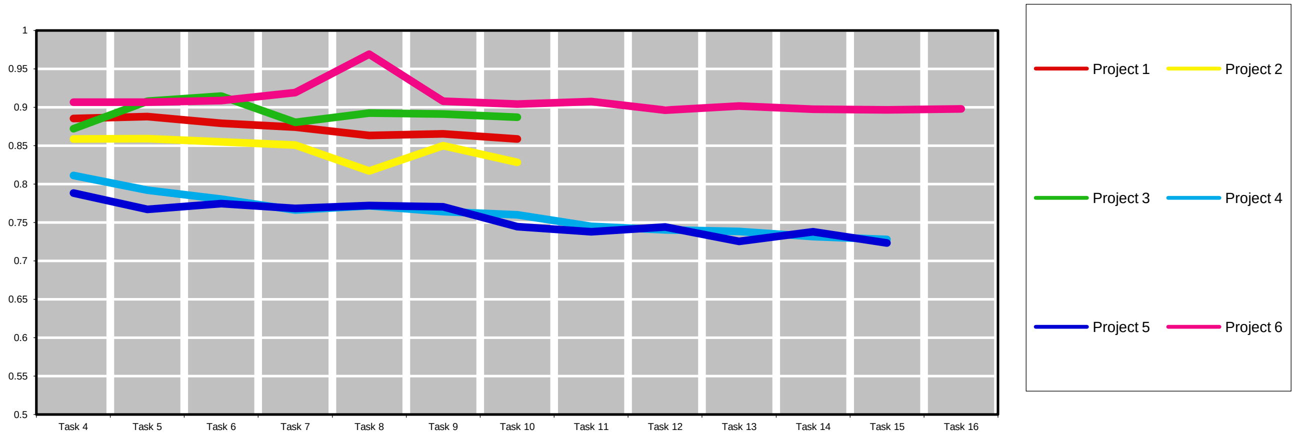


Figure 4a: Different measures throughout conjoint questionnaires – Percent Certainty

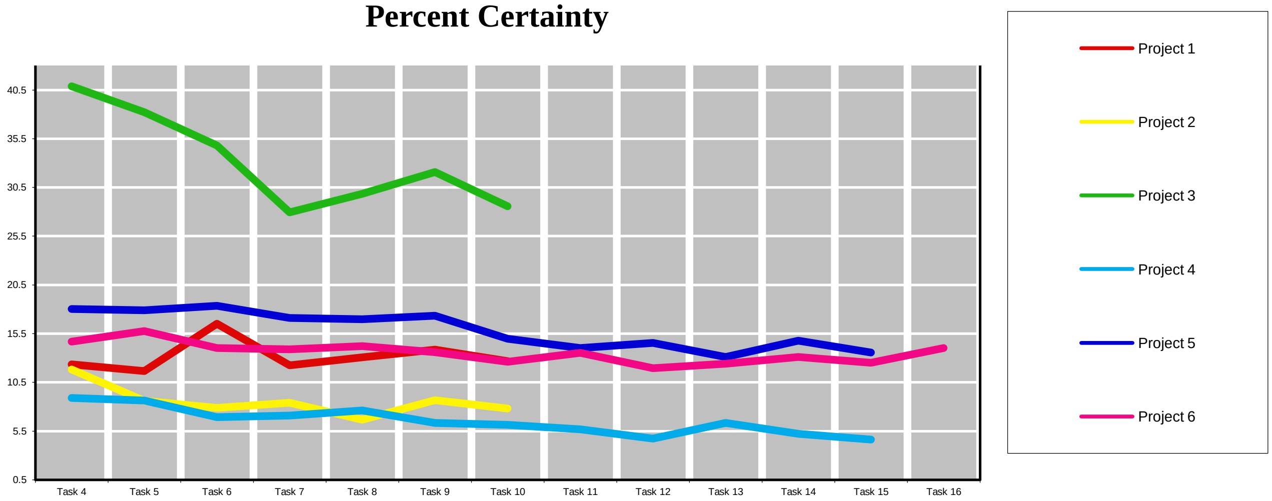


Figure 4b: Different measures throughout conjoint questionnaires – Root Likelihood (RLH)

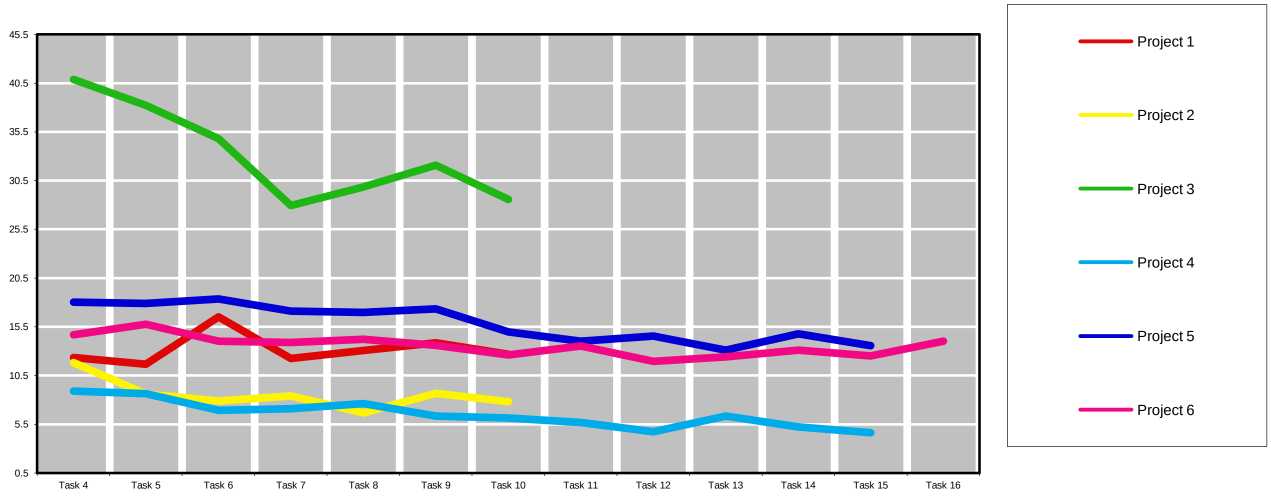


Figure 4c: Different measures throughout conjoint questionnaires – Variance

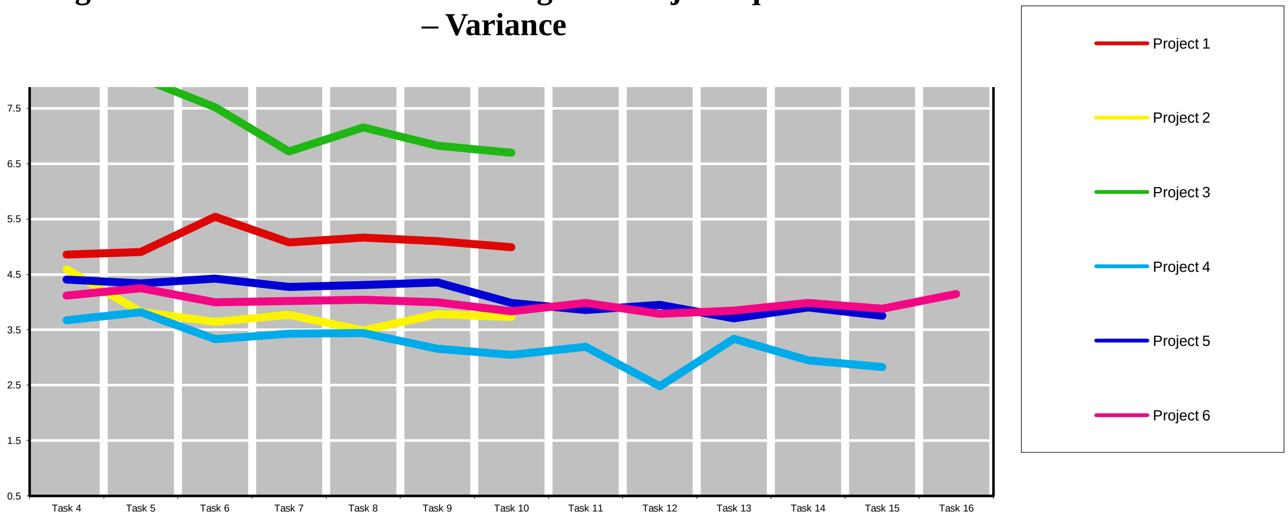
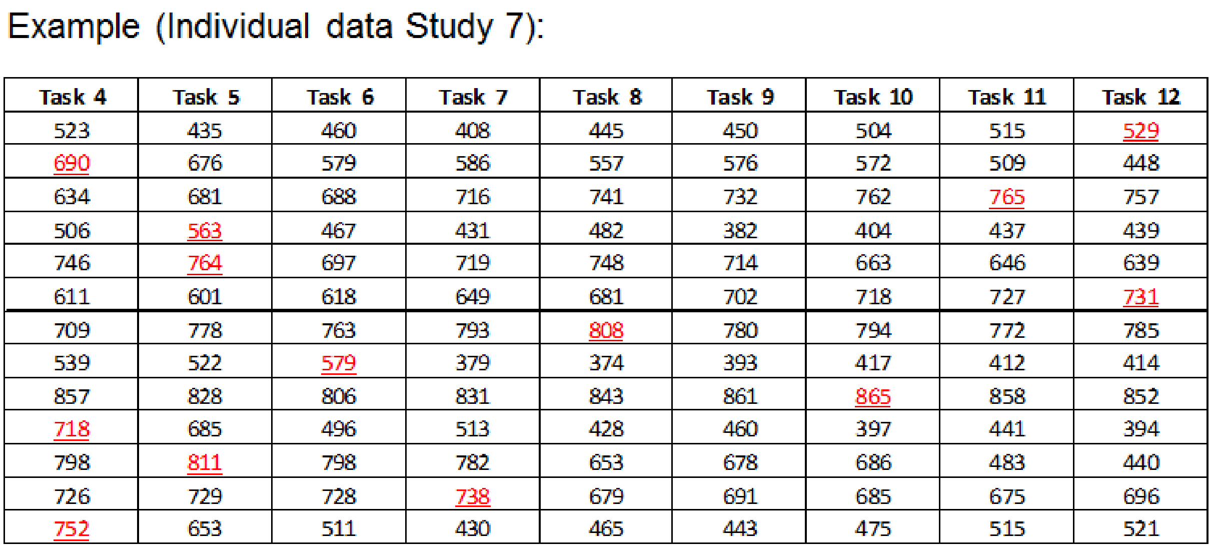


Figure 4d: Different measures throughout conjoint questionnaires – Parameter RMS

On an aggregated level we see that from 4th to last choice task all four measures tend to the same values.

Would that mean that four choice tasks are enough? This is certainly not the case. These measures are only sample averages and don't provide any information about the heterogeneity of respondents.

The individual RLH results showed a different picture, consistently in all the 12 studies. Here as an example is a cutout of individual data from one of our studies:



Example (Individual data Study 7):

Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12
523	435	460	408	445	450	504	515	529
690	676	579	586	557	576	572	509	448
634	681	688	716	741	732	762	765	757
506	563	467	431	482	382	404	437	439
746	764	697	719	748	714	663	646	639
611	601	618	649	681	702	718	727	731
709	778	763	793	808	780	794	772	785
539	522	579	379	374	393	417	412	414
857	828	806	831	843	861	865	858	852
718	685	496	513	428	460	397	441	394
798	811	798	782	653	678	686	483	440
726	729	728	738	679	691	685	675	696
752	653	511	430	465	443	475	515	521

Figure 5: Typical individual RLH results

The highest individual RLHs (marked in red and underlined) are reached at a different number of choice tasks for each respondent. Also in a closer look at the data, where we counted how often the maximum individual RLH was reached at each number of tasks, we found no clear pattern.

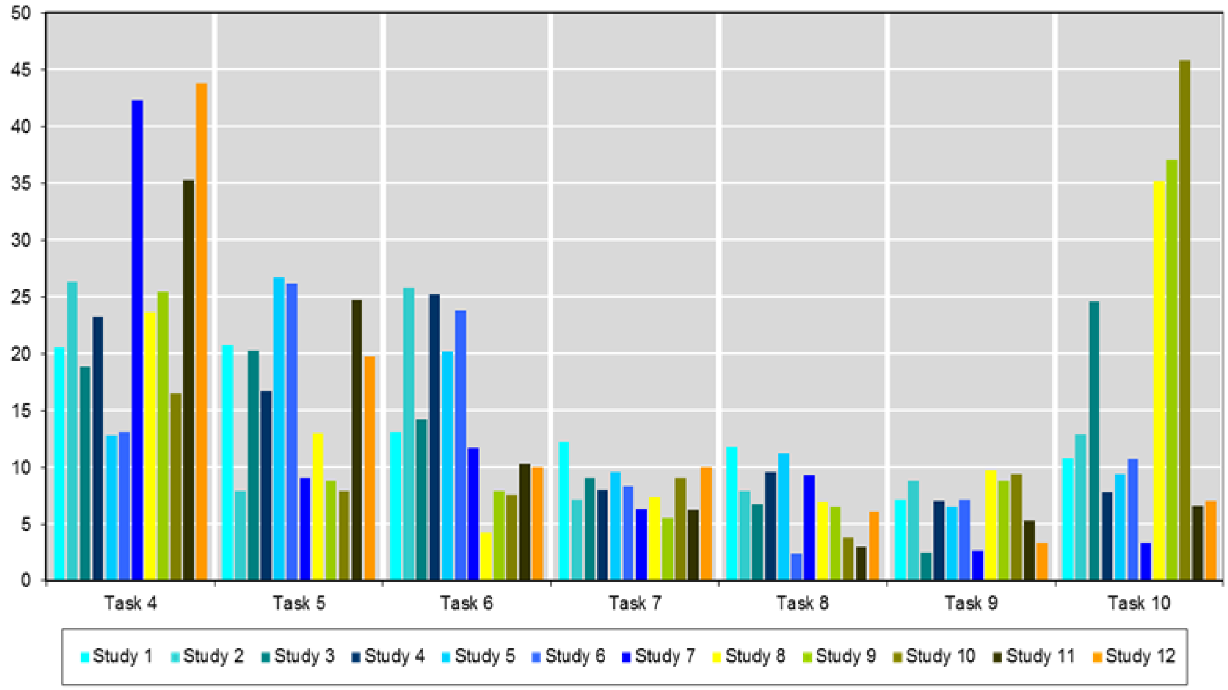


Figure 6: Highest individual RLH per task in %

Knowing that the RLH measure is dependent on the amount of information we use for the HB estimation, RLH might not be the best measure for comparisons across different numbers of tasks. RLH inherently decreases with the number of choice tasks used for deriving the part worth utilities (simply because having more tasks makes overfitting less possible). Therefore we repeated our analysis by creating an index taking the aggregated difference in RLH from one choice task to the next into account when identifying individual decreases in RLH. However, the results were the same: much individual variation in when the highest RLH was achieved.

Impact on Share Prediction

Our next question is whether there is a difference in share prediction with different numbers of choice tasks. Therefore we created datasets with increasing number of choice tasks and ran simulations against hold-out tasks. The MAE (mean absolute error) tends to decline with an increasing number of choice tasks. However, as Figure 7 shows, in some studies there is a stagnation after 6 or 7 choice tasks; in one study the MAE was even increasing after that. Overall there we can say that not all projects benefit from additional choice tasks.

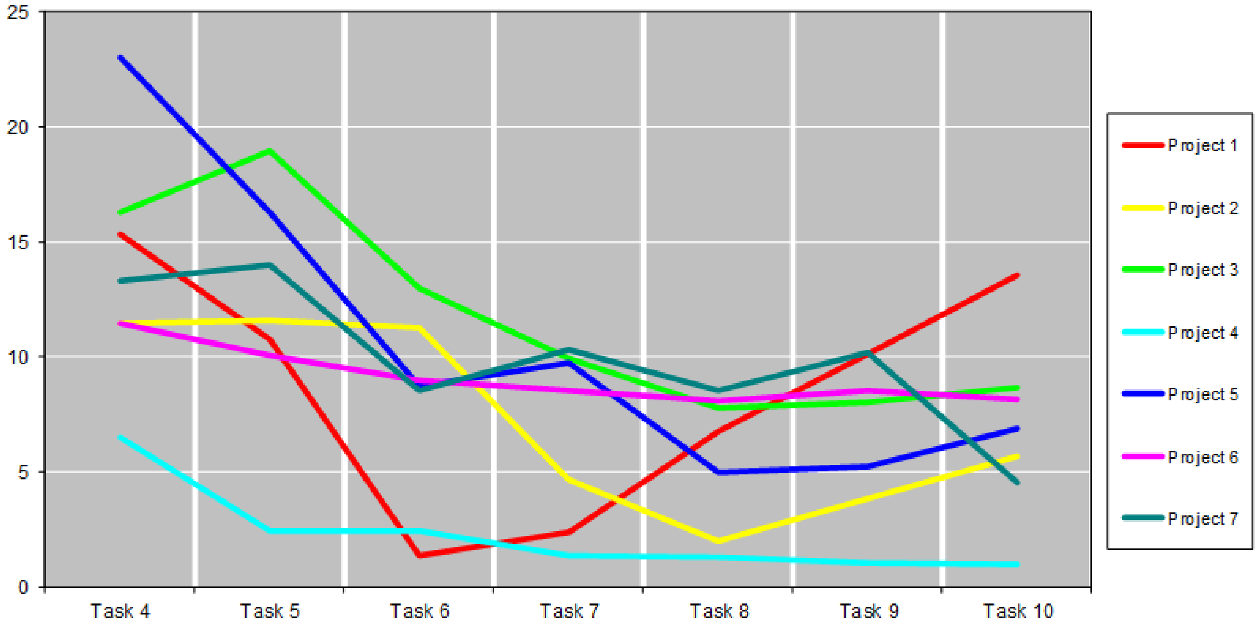


Figure 7: MAE on hold-out tasks by task

Also the position of the choice task after which a significant individual RLH decrease could be measured was widely distributed. However, it seemed that beginning with task 7 this effect can be measured more and more often:

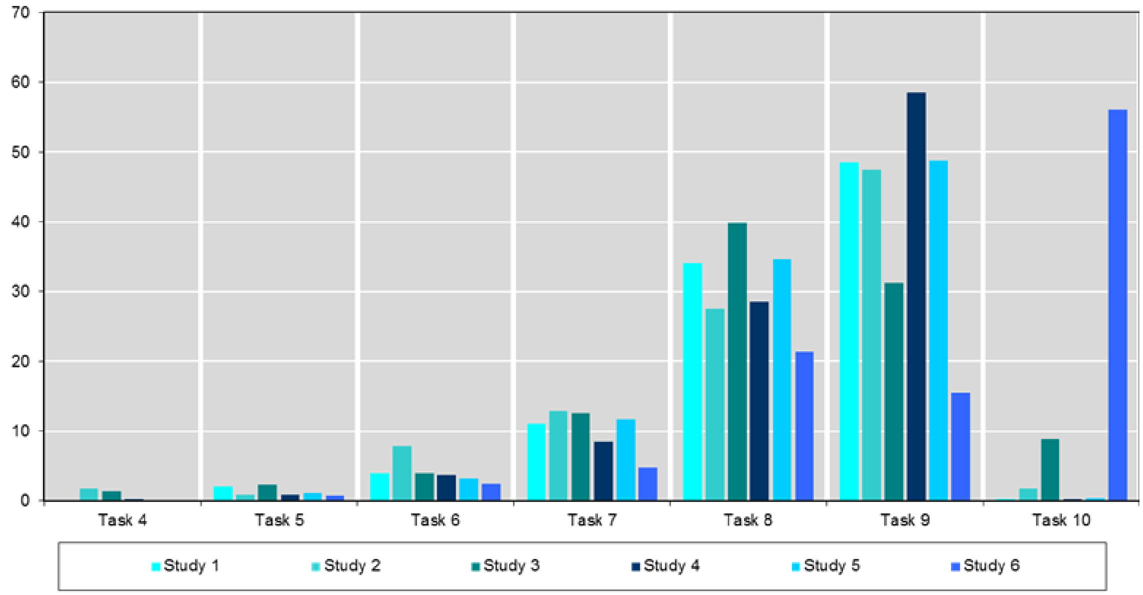
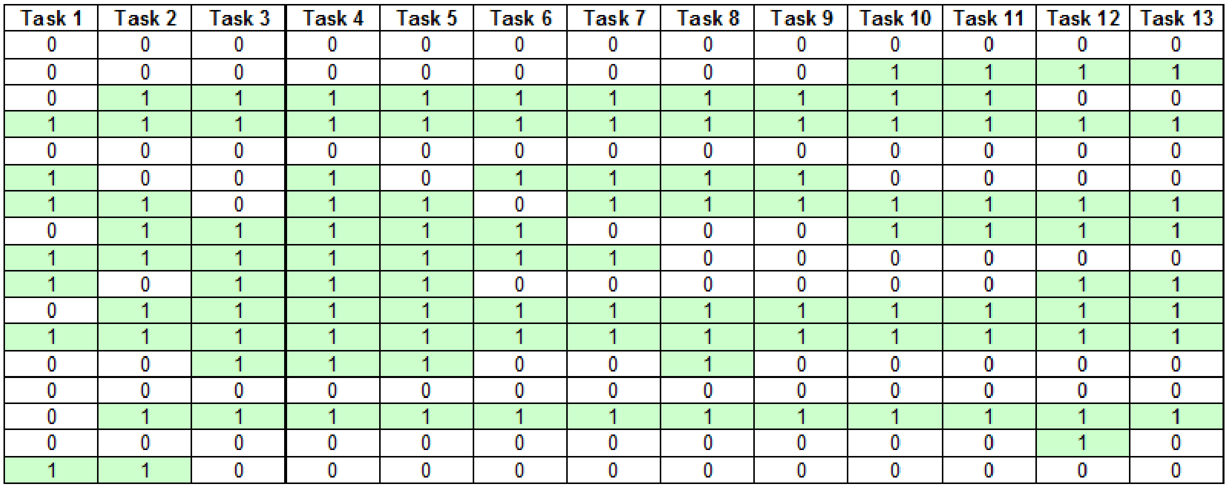


Figure 8: significant individual RLH decrease by choice task

Again using datasets with increasing number of choice tasks we calculated individual hit rates. These are also distributed all over the tasks without a clear indication of an overall choice task threshold.



Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12	Task 13
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	1	1	1
0	1	1	1	1	1	1	1	1	1	1	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	1	1	1	1	0	0	0	0
1	1	0	1	1	0	1	1	1	1	1	1	1
0	1	1	1	1	1	0	0	0	1	1	1	1
1	1	1	1	1	1	1	0	0	0	0	0	0
1	0	1	1	1	0	0	0	0	0	0	1	1
0	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1
0	0	1	1	1	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	1	0
1	1	0	0	0	0	0	0	0	0	0	0	0

Figure 9: Example data: individual hit rates by choice task

If we accumulate the respondents with at least one hit and don't take into account the later tasks of the respondent, we reach a high level of hits after only a few choice tasks:

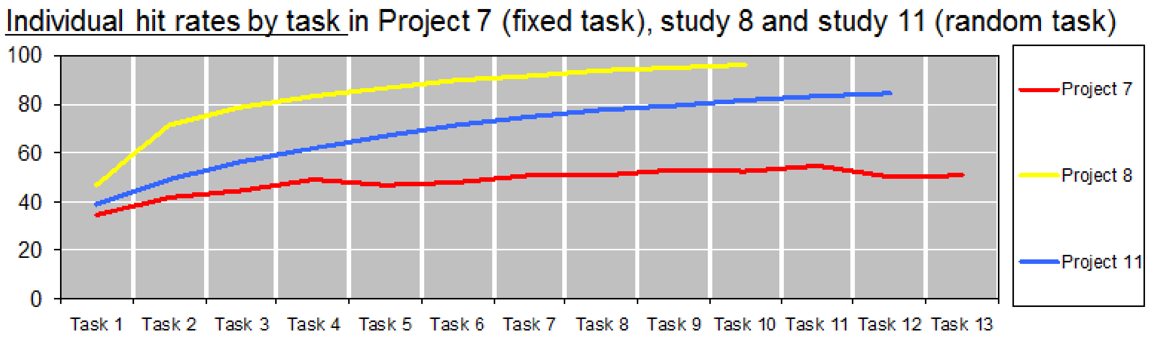


Figure 10: Accumulated individual hit rates by task (fixed holdout task used for study 7; for all other projects the first task which is not used for estimation (n+1) is used as holdout)

Looking at the three studies for the task in which an individual hit has first been reached we see after only a few choice tasks there is only very slow improvement.

Individual Choice Task Optimization

Based on our conclusions so far we optimized our data sets: We used only those choice tasks up to the RLH maximum on an individual level (meaning different individual interview lengths within the different studies). Applying this rule we could eliminate about 38% of all choice tasks. Were these choice tasks necessary at all?

Looking at total answer time and aggregated hit rates we found quite similar results for the ICT (individual choice task threshold) optimized data sets. In reality this would have led to a time saving potential of about 35.4%.

		Study1	Study2	Study3	Study4	Study5	Study6	Study7	Study8
All Choice Tasks	Choice Tasks	6.076	2.040	12.600	6.900	7.695	12.835	3.900	2.120
	Answer Time (hours)	47,3	9,6	66,5	26,8	27,8	71,3	17,5	4,3
	Hit Rate (percent)	**69,2**	**73,6**	**50,7**	**96,3**	**78,2**	**85,4**	**50,7**	**96,3**
ICT Optimization	Choice Tasks	3.957	1.499	7.823	3.879	4.957	7.953	1.957	1.579
	Answer Time (hours)	30,8	7,1	41,3	15,8	17,9	44,2	10,6	3,0
	Hit Rate (percent)	**68,0**	**68,4**	**48,3**	**85,6**	**64,2**	**78,1**	**48,3**	**85,6**
	Time Saving Potential	**35,9%**	**26,0%**	**37,9%**	**41,1%**	**35,6%**	**38,0%**	**39,4%**	**30,2%**

Figure 11: Effect of ICT optimization

The above results lead to the conclusion that Individual Choice Task Thresholds (ICT) exist. Unfortunately, since they are not correlated with the answer times, they can't simply be detected by time control. Using an individual number of choice tasks according to the individual choice task threshold does not end up with large differences in the hit rates. The small decrease we can see in all of our eight studies does not seem to change the interpretation of the results in any way and therefore we are focusing more on the costs and interview length. An individual number of choice tasks shortens the interview significantly which would have several positive effects such as:

- lower field costs (time is money)
- less frustration of respondents who are either challenged or bored
- keeping the panels open for conjoint exercises

However, there is another interesting aspect: while the hit rates are quite similar between the full and ICT optimized data sets, we observed important differences in the results:

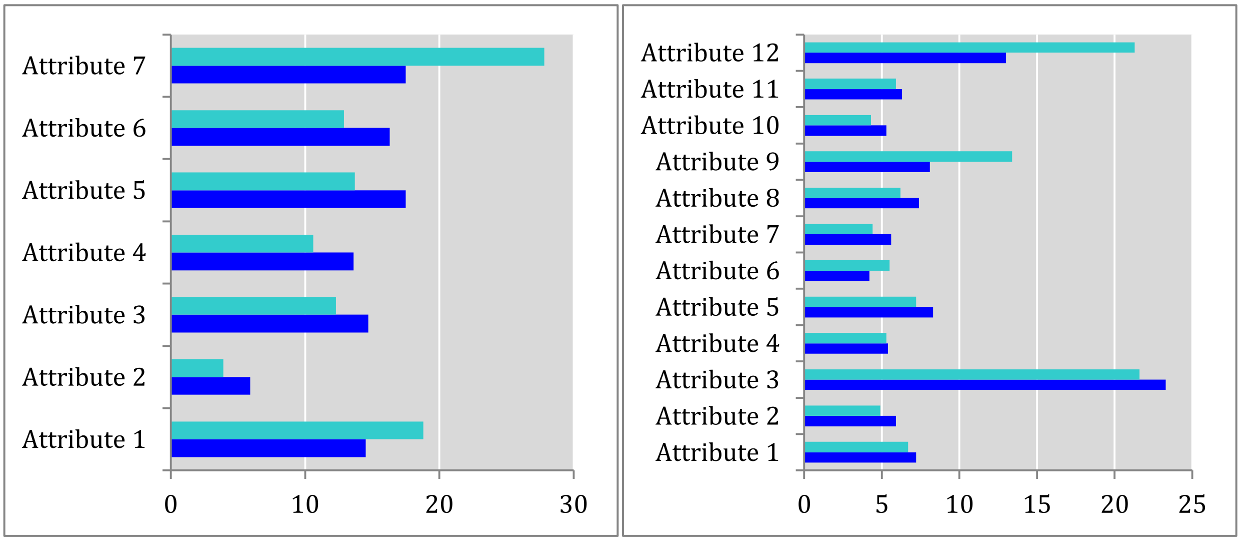


Figure 12: Differences in Average Importances between all tasks and ICT optimized datasets

It seems that some attributes become more important during those choice tasks which do not contribute to better individual results and were therefore excluded in the ICT optimization. This definitely leads to wrong interpretation of the impact of certain attributes.

Comparison of the Posterior Distribution

As shown in the previous section the individual hit rate sometimes increases, while share prediction results show a higher MAE. Simply looking at the aggregated results, it seems counter- intuitive that if one gets better hit rates with more choice tasks the MAE is getting worse.

A closer look into the data structure is necessary to see how this effect happens. The posterior distributions of the HB estimates and their standard deviations can provide more information. If the ICT really exists, a decreasing number of attributes with part worth utilities significantly different from zero must be found when increasing the number of choice tasks in the estimation. The reason for this effect is a simplification strategy, namely focusing on a lower number of attributes the more choice tasks we ask. If it's possible to show this decrease in the number of part worth utilities significantly different from zero, we could conclude that respondents concentrate on only a small number of attributes when answering a larger number of choice tasks.

To show this effect we calculated[9] the number of attributes with significantly non-zero part worth utilities for each respondent, after analyzing with each additional choice task. Comparing the results of the twelve conjoint studies we found that the more choice tasks we use in the estimation, the more respondents take into account a smaller number of attributes when answering.

The more choice tasks we show, the more respondents consider only one or two attributes before making their choice. In contrast, at the beginning of the choice exercise, respondents show a more complex answering behavior and take into account a higher number of attributes when making their choice.

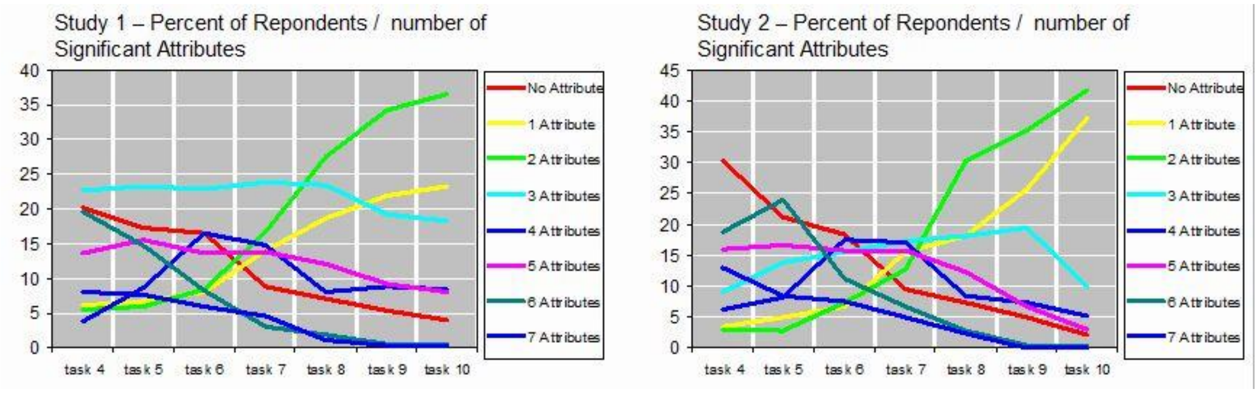


Figure 13: Number of attributes taken into account / % of respondents

The tendency toward taking fewer attributes into account before selecting a concept in the tasks could be observed in all twelve analyzed studies. The number of attributes in the choice experiment seems to have a much smaller influence on using the simplification strategy than does the number of choice tasks the respondent has to answer. By far the largest respondent group takes only two attributes, brand and price, into account in the later tasks. Only very small numbers of respondents focus on more than four attributes in later choices. Especially in studies with many different feature attributes, respondents tend to simplify very early. In Brand-Price only CBC we can show a very fast convergence into two different simplification strategies, one group taking only brand, and the other only price, into account before making their choice.

[9] We used the individual respondent beta draws from HB and report attributes that included levels where 95% or more draws have the same sign. This shows how many attributes the respondents takes into account when making their choices.

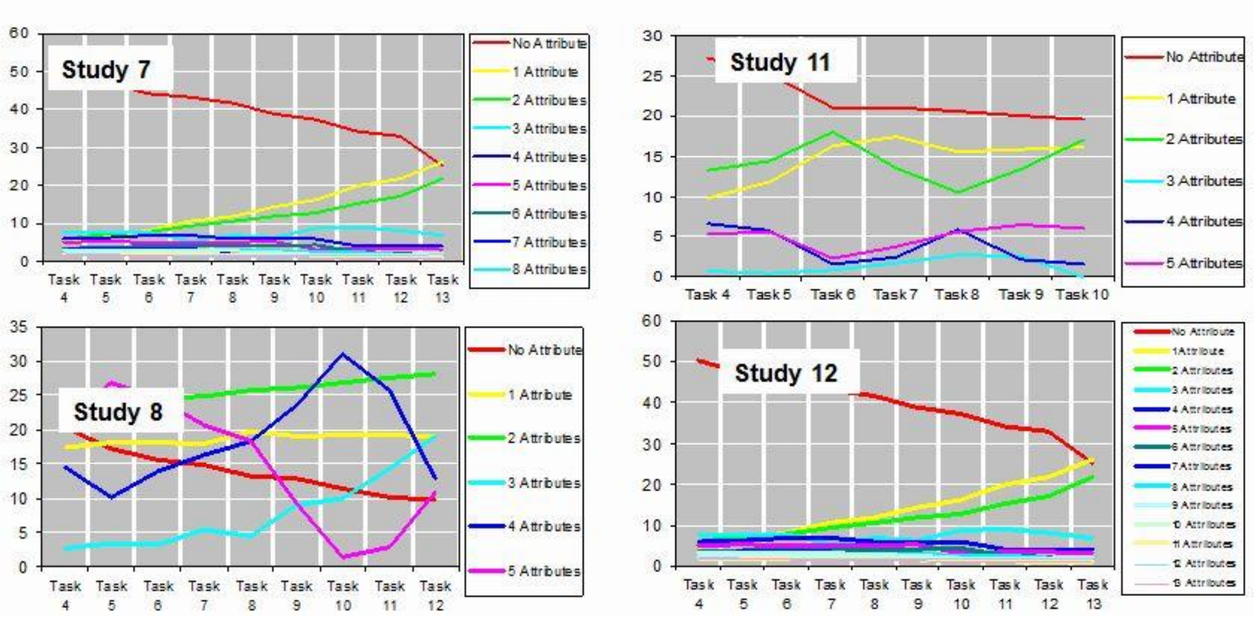


Figure 14: Number of attributes taken into account / in % of respondents

The decreasing number of attributes that are taken into account seems to be at least one of the reasons for increasing hit rates when asking more tasks, but increasing MAE in our share predictions.

A slight tendency towards higher MAE for high involvement products underlines the finding, because it’s plausible, that for high involvement goods the simplification doesn’t take place with the same intensity that it does for low involvement goods. In the case of low involvement, one can argue that the simplification makes a lot more sense because it reflects reality. People don’t look at all the tiny differences between products if they are not really interested in this category.

The findings from our twelve studies show that—in at least ten of them—simplification confounds the real decision process and leads to incorrect share predictions. A simple example illustrates these findings. If respondents focus mainly on brand and price after a certain number of choice tasks, then their hit rate on a holdout or random task in a late position is improved because the choice behavior is much easier to predict. However, if in reality their real purchase decision is based on more attributes than brand and price, then the prediction of market shares based on their part worth utilities is inferior. That is the reason why better hit rates don’t always result in better share prediction.

IS THIS A NEW PROBLEM?

Maybe it is. Today we conduct most interviews in online panels and this environment has changed during the last decade dramatically.

We observe a concentration of panels. Many mergers and acquisitions take place in this field. Often, high quality panels are taken over by ones with lower quality, resulting in one lower quality panel.

More and more conjoint exercises are conducted around the world and therefore panel respondents are more often used to answering conjoint studies. In the past, most of the respondents were seeing a conjoint exercise for the first time when answering and didn't really know what the next screen would bring up. Now, panelists more and more know the flow, especially of the widely used classical CBC exercises, and know how they can speed up when answering without being identified as bad interviewees. This comes along with the fact that most of the panelists only respond for points, miles or other incentives. So from their motivation, it's best to finish the exercises as fast and efficiently as possible to earn incentives as efficiently as possible. In many of the feedback mails we received from panelists, after participating in a conjoint interview in the last year, we could read that the more engaging conjoint exercises are seen as a higher burden in earning points than simple scaled questions are. So panel providers are asked to offer higher incentives for these burdensome exercises in order to motivate panelists to take part in future conjoint studies. Many of the respondents say they will stop future interviews with conjoint exercises if the incentives are not higher compared to studies with conventional questions only.

Furthermore, we see in feedback mails that many of the panelists are less motivated and not willing to answer boring choice tasks any longer. The monotonous repetition of a large number of very similar-looking questions is often their main complaint. So from this feedback we could, conclude that we are forced to program more interesting choice exercises in future and make our studies more appealing to our panelists.

Nevertheless neither higher incentives nor fancy programming of the survey can protect us from the "speed up" behavior and the simplification strategies of the professionals in the panel, because more points can be earned in an efficient way by again speeding up the exercise through simplification.

TO PUT IT IN A NUTSHELL

The analysis conducted for this paper clearly shows that Individual Choice Task Thresholds (ICT) really exist. In our twelve analyzed studies we could see that respondents have a very diverse answering behavior and individually different choice task thresholds.

Unfortunately there is no simple rule for detecting the ICT during the online interview. Neither the absolute time a respondent needs to answer a question, nor the change in time needed for a single answer, provides useful information about the answering quality and behavior of a respondent. A speed-up on the one hand could be a sign of a simplification or, on the other hand, a learning process enabling the respondent to make decisions much faster. For this reason, analyzing the time stamps from questions alone doesn't really help. Only a more complex analysis taking more advanced measures into account could provide us with the information we need to decide if a respondent has reached her ICT.

The key findings of our work are:

Less is more: For a large number of respondents we gain better or equal hit-rates and share predictions when we use only a smaller number of choice tasks in order to avoid simplification.
More is dangerous: The analysis of the individual posterior distributions shows that a large number of respondents tend to simplify their answers in later choice tasks. In earlier choice tasks, one observes higher number of attributes with significantly non-zero utility values than in later ones.
More is expensive: Optimizing the number of choice tasks could reduce the total number of tasks without a significant quality decrease, thus saving about a third of interview time (for the conjoint part) and avoiding interviewee annoyance.

CONCLUSION

A better monitoring of individual answering behavior during the interview in order to avoid exceeding the ICT will be essential in the future.

A combination of monitoring the answering time and a maximum likelihood estimation predicting the next choice of a concept, based on respondents' previous answers could be an idea for future research. Another new approach could be a surveying system with individual number of choice tasks i.e. derived by measuring individual RLH against the aggregated RLH (Indexed LH) after each choice task or by comparing the posterior distributions from one task to the next. The system should stop the interview if there is no further improvement in RLH or simplification strategy or consistently lower measures for interview quality are detected. As computational power is increasing from year to year it seems possible to realize such online computations during online interviews in the near future (Kurz, Sikorski; 2011).

Another interesting approach that could be developed would be the combination of higher incentives for panelists and a more appealing conjoint programming in combination with an adaptive design algorithm. In the work of Toubia, Hauser and Garcia (2007) and Yu, Goos and Vandebroek (2011), we find new ideas for adaptive design algorithms with the aim to produce choice tasks with higher utility balance for each individual respondent. Perhaps combining higher utility balance with criteria that stop the interview when reaching a certain ICT would provide a solution. Calculating the maximum likelihood after each choice task and forecasting the answer to the next choice task would allow stopping the interview after this forecast produces a hit two or three times. One could then assume that enough information from this respondent is collected. This technique could avoid respondents simply clicking through conjoint questionnaires in order to earn the points. To be sure, this exercise is more burdensome, because choice tasks with higher utility balance are much harder to answer for respondents. But such adaptive interviews can be rewarded with using fewer choice tasks when answer quality is high and respondents can earn the same amount of incentives with an individually different length of the interview. Such a combination of adaptive algorithms and higher incentives could be most beneficial, and avoid only paying the higher incentives without getting improved quality for the conjoint interviews. In cases when no alternative-specific designs are needed[10] a solution can also be ACBC, because ACBC's more appealing interview leads to more respondent

[10] At the time of writing, Sawtooth Software's current version of ACBC software cannot do alternative-specific designs. Sawtooth Software says that soon this will be available.

engagement. ACBC uses different stages with differently formatted choice questions, and in the minds of most respondents is not as boring as CBC tasks. Maybe interspersing non-conjoint questions and transitional text-only pages throughout the CBC Tasks could help a little to reduce the monotony and avoid burning out respondents during the CBC exercise.

References

Johnson, R., Orme, B., (1996) "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" in Sawtooth Software, Inc. Technical Papers.

Kurz, P., Binner, S., (2010) "Added Value through Covariates in HB," in Proceedings of the 15th Sawtooth Software Conference, pp.269-282.

Kurz, P., Sikorski, A.,(2011) "Application Specific Communication Stack for Computationally Intensive Market Research Internet Information System" in: Proceedings of the 14th International Conference on Business Information Systems, pp. 207-217.

Toubia, O., Hauser, J. R., Garcia, R., (2007) "Probabilistic polyhedral methods for adaptive choice based conjoint analysis: Theory and application" in Marketing Science, 26(5), pp. 596–610.

Yu, J., Goos, P., Vandebroek, M., (2011) "Individually adapted sequential Bayesian conjoint-choice designs in the presence of consumer heterogeneity" in International Journal of Research in Marketing, vol. 28, pp. 378-388.

TAKING NOTHING SERIOUSLY
OR
"MUCH ADO ABOUT NOTHING"

KEVIN D. KARTY
BIN YU
AFFINNOVA

ABSTRACT

Most conjoint and choice based applications deploy a 'None' concept or button of some sort on each choice set. Respondents are notorious for ignoring or underutilizing the 'None' selection, resulting in model estimates that understate the likelihood of not purchasing or using a product or service in the real world. Existing techniques, like the dual response method, can somewhat reduce this bias but do not eliminate it. This paper proposes and demonstrates new methods that dramatically increase respondent usage of the 'None' option, and offers a case study demonstrating strong improvements in external validity.

BACKGROUND

When conducting conjoint and discrete choice models, researchers are frequently interested in understanding the likelihood that respondents would not buy or use anything in the set of items (or bundles of features) that is simulated. Depending on the context of the study, a researcher may hope to measure how many consumers will choose not to purchase anything in a category, not to apply for a credit card, not to download a software application, and so forth. In these situations, the share of respondents which select each item from a set of allowed items can be less important than the share of respondents who select nothing at all.

Choice models that attempt to predict how many consumers will buy nothing rely heavily on respondents' selection of a 'None' option in the choice exercise itself. Quite simply, any bias or tendency that causes respondents to choose a 'None' option less frequently than they would choose not to purchase something in real life will cause a model to over-predict the number of consumers who are likely to buy a product. This over-prediction can be quite dramatic in almost any environment.

In monadic (Likert scale) questions, overstatement of purchase interest is a common problem, which has led to the development of 'normative' calibration factors to down-weight aggregate claimed purchase to better predict likelihood of purchase.[11] These normative factors are often category, context, culture, and country specific.[12]

Overstatement of purchase intent (or, equivalently, under-selection of the 'None' option) in choice exercises has also been recognized as an issue for many years. One of the more comprehensive reviews of the use of usage of the 'None' option in choice studies was conducted

[11] Kevin J. Clancy, Peter C. Krieg, and Marianne McGarry Wolf. *Marketing New Products Successfully: Using Simulated Test Market Technology*, (Lanham, MD: Lexington Books, 2006).

[12] The most popular monadic-based market sales projection systems, BASES, maintains a database of tens of thousands of product launches.

by Pinnell, Fridley, and Ulrich in 2006.[13] Summarizing work from Huber and Pinnell (1994) and Johnson and Orme (1996), they report relatively low overall 'None' usage and significant but relatively small variation in none usage as the features changed. Much of the related work they review assesses the risk of decision avoidance, with conflicting results that may indicate some relationship to the specifics of individual studies. This line of academic work ultimately led to the development of the 'Dual Response' approach.

The Dual Response approach, in which respondents are first asked to select an item from a choice task and then to indicate if they would actually buy it, was first detailed in 2002.[14] Additional research conducted up through 2006 and incorporation of the technique into Sawtooth Software's core CBC offering has since increased its popularity.[15] Although it was commonly observed that Dual Response techniques tended to increase the likelihood of none selection, this has not been the primary focus of investigations into the technique. Rather, academic work has largely focused on the impact of Dual Response on improving the precision of estimated utilities by gathering relative preference information even in choice sets where respondents selected the 'None' option, as well as the impact of Dual Response on respondent vulnerability to specific context effects.

Subsequent to Sawtooth Software's incorporation of Dual Response into its CBC offering, research on the issue slowed. Investigations by Pinnell, Fridley, and Ulrich (above) offered some additional insights, however. Notably, they consider alternatives to the 2-point scale (yes/no) follow-up question, including a 5-point scale question which is variously asked in an interwoven manner (following each choice task) or at the end of all choice tasks. The 5-point scale follow-up appeared to increase prediction of the 'No Buy' option. They also observed that asking a BYO (build your own) task prior to the choice task section of the study increases usage of the 'No Buy' option slightly, but were unable to draw strong conclusions about whether changes to respondent behavior over the course of the study resulted from respondent learning or from fatigue.

Although much work has been invested in understanding respondents' usage of the 'None' option, little of that work has focused on the empirical validity of different implementations. Specifically, which techniques to solicit respondent preference for the 'No Buy' or 'None' option have the best chance of predicting actual respondent likelihood to not purchase a product in market?

OUR EXPERIENCE WITH 'NONE'

Like many researchers, we have consistently observed that estimates of 'None' share derived from using the 'None of the Above' option in choice tasks yield unreasonably low projections of the likelihood that consumers will choose not to buy the product or service being studied, even after adjusting for factors such as awareness and distribution. We provide some examples below.

[13] Jon Pinnell, Lisa Fridley, and Meagan Ulrich "Alternative Methods to Measure Changes in Market Size in Discrete Choice Modeling" Asia Pacific Quantitative Methods in Marketing Conference 2006

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

[15] Uldry, Pierre; Valerie Severin and Chris Diener. "Using A Dual Response Framework in Choice Modeling," 2002 AMA Advanced Research Techniques Forum, Vail, Colorado. Also: Chris Diener, Bryan Orme and Dan Yardley. "Dual Response 'None' Approaches: Theory and Practice. Proceedings of the Sawtooth Software Conference, March 2006.

In Figure 1, we summarize the selection rates for the 'None' option in an older conjoint study on store layouts and formats. With 25 choice sets, the study was somewhat burdensome to respondents. The blue line indicates the count of the number of people who selected 'None' a given number of times. Thus, 500 respondents never selected 'None', and about 100 selected 'None' on only one of the 25 choice tasks. The orange line indicates the cumulative percentage of that selected 'None' a certain number of times or fewer. Thus, 60% of respondents selected 'None' on 4 or fewer of the 25 choice tasks.

Figure 2 shows the share of respondents who selected 'None' on each choice task. Our primary observation is that this share is empirically quite low. At face value, it is an unrealistically low projection of how many respondents might choose not to shop at one of the four store concepts displayed in a choice task. We also note that the trend increases such that 15% of selections are 'None' for the first few choice tasks but 25% are 'None' for the last few choice tasks. Although this trend does not consistently appear in all studies, it reflects concerns similar to those expressed by prior researchers about learning effects, sincerity, and fatigue.

Figure 1

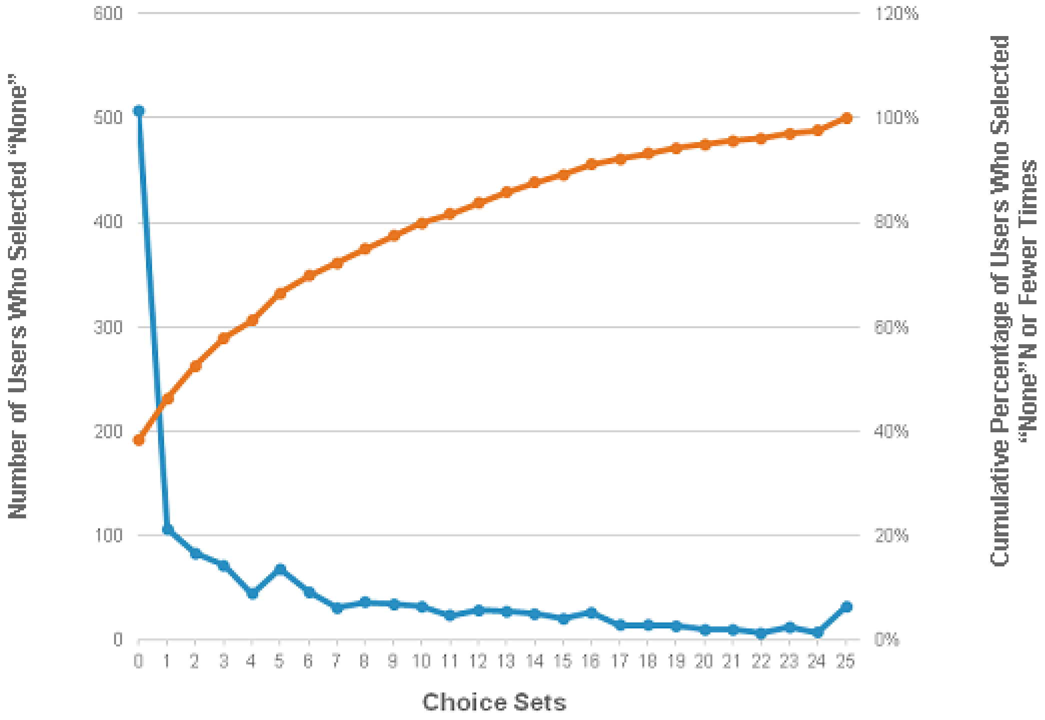
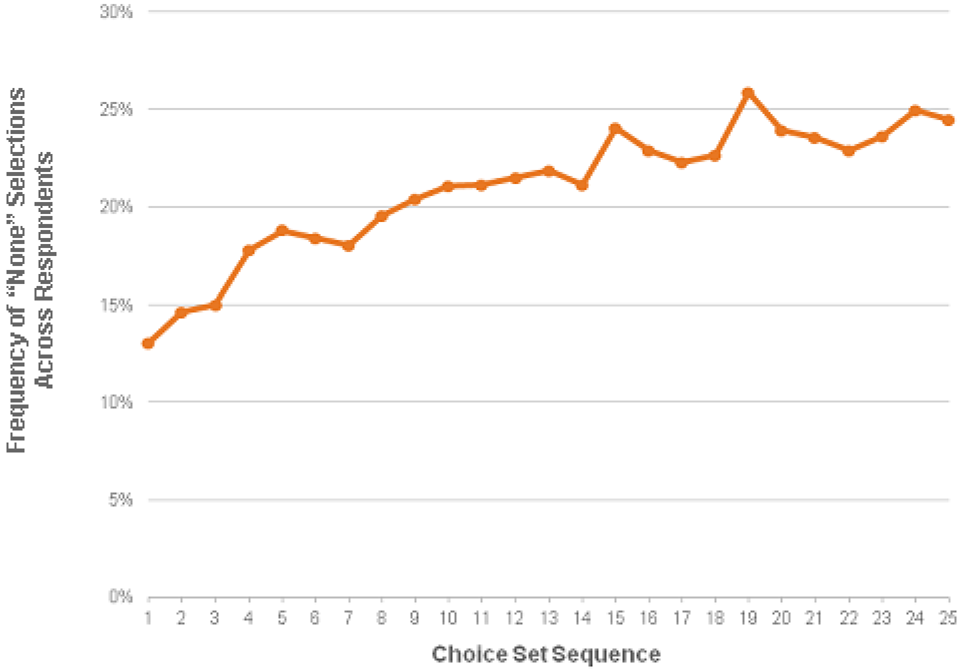


Figure 2



Initial Attempts to Improve 'None': Offering Multiple 'None' Options

One of the key theoretical problems with the 'None' option is that it is not specified. That is, the 'None' option is a generic, non-specific external good. As such, the most likely interpretation of the 'None' option is the intention to keep the money, however we know from a large literature that consumers tend to overstate willingness to pay in choice studies (and virtually every other pricing method).[16] Various improvements to survey design have been suggested to improve willingness to pay estimates (and the realism of choices), including a great deal of work on incentive compatibility.[17] One simple approach to improving respondent usage of the 'None' option was to offer a specific interpretation to the 'None' option to help anchor respondent understanding.

Our first attempts to implement this involved offering the respondent multiple 'None' buttons, each representing a different alternative in an adjacent category. In the attempt described below, the client's goal was to optimize the product assortment for a category in which there was no direct competition, but there was a great deal of indirect competition. The category was a unique snacking category with medium to high awareness but low category penetration. We specified a basic 'None' option and four additional 'None' options representing adjacent alternative categories that consumers might substitute for the product. Our goal was to increase respondent usage of the 'None' option to reduce preference share saturation (i.e. the tendency of preference share to rapidly approach 100% with just a small number of SKUs in the lineup).

In the data shown below, this initial attempt to increase 'None' usage did not appear successful. In Figure 3, we observe that over 70% of respondents selected 'None' on 4 or fewer choice tasks out of 20, and average 'None' usage was between 15% and 20% throughout the study (with no clear trend line, as shown in Figure 4). The simulated purchase rate for the current line of products was 84% when including all five 'None' options, even though actual category penetration was well under 20%. This was an improvement over simulating share vs. the single generic 'None' option (which yielded a 90% acceptance rate), but we do not know whether we would have seen increased relative usage of the generic 'None' in the absence of the other four category-alternative 'None' options.

[16] For example: Harrison, Glenn W. and Elisabet E. Rustrom. 2005. "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods." In Charles R. Plott and Vernon L. Smith (eds.), *Handbook of Experimental Economics Results.* New York: Elsevier Press.

[17] For example: Songting Dong, Min Ding, and Joel Huber. "A simple mechanism to incentive-align conjoint experiments" International Journal of Research in Marketing. Vol. 27. 2010:25-35.

Figure 3

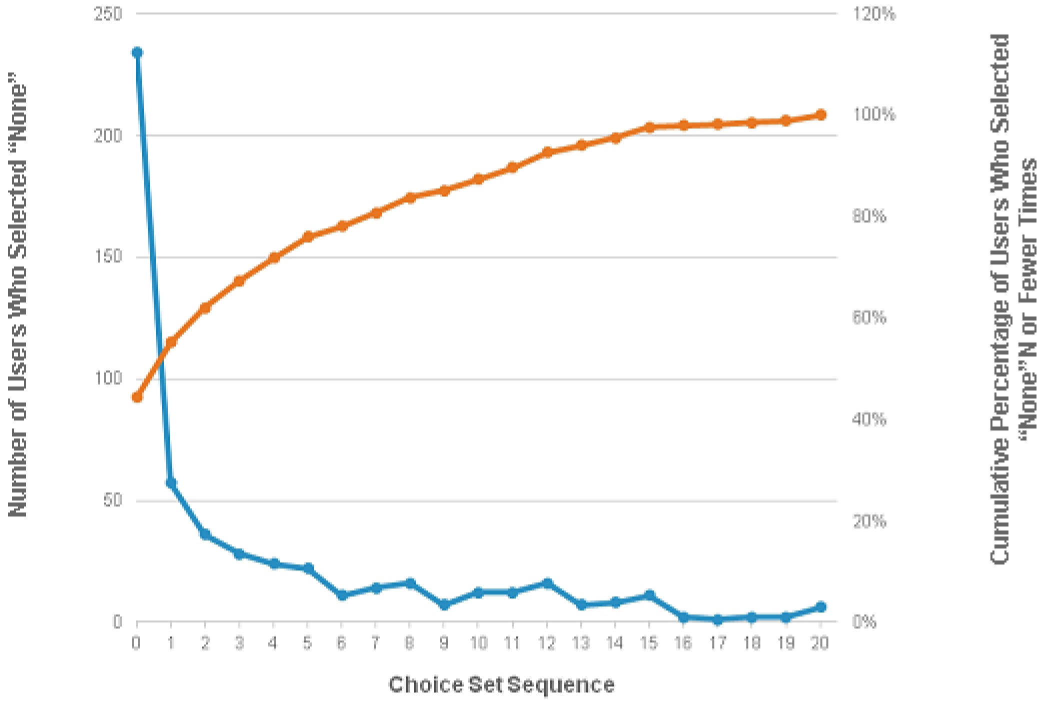
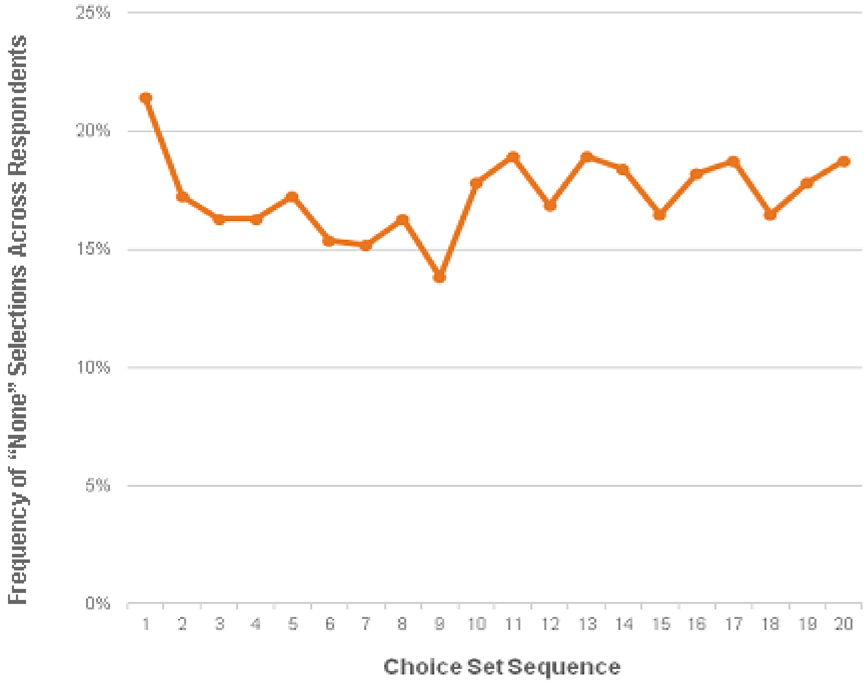


Figure 4



INITIAL ATTEMPTS TO IMPROVE 'NONE': EXPANDING DUAL RESPONSE TO A 5-POINT SCALE

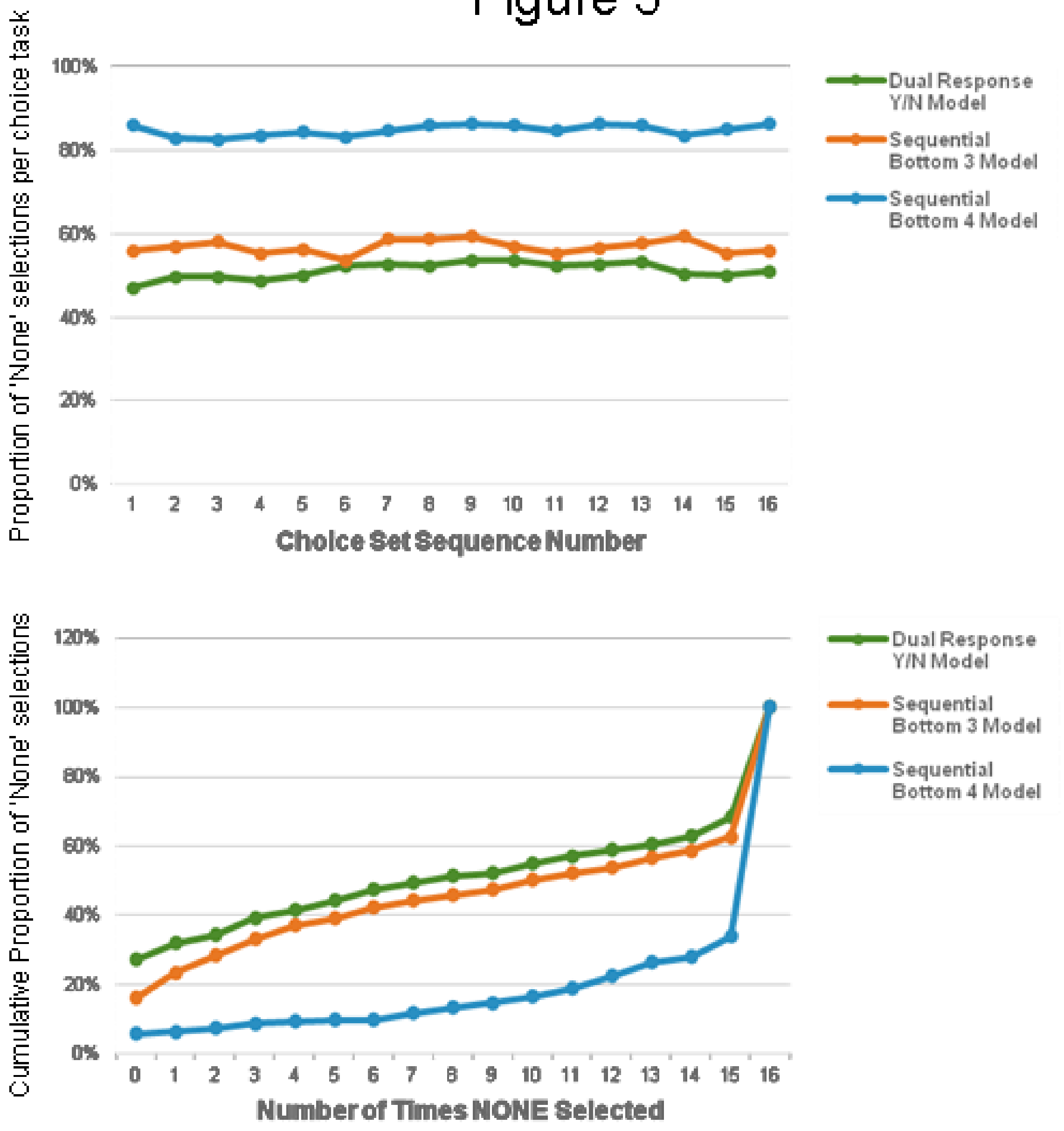
Recognizing the relative success of Dual Response in increasing 'None' selection, we attempted to expand on this success by expanding the scale. Expanding the scale is a simple method to enhance the amount of discrimination in a question. This approach is consistent with work described above by Pinnell et al. (2006), although we were unfamiliar with their work at the time. Similar research in this area was also presented in our earlier work in a different context.[18]

In the Figures below, we show the results from a study conducted on a new financial services product. The client had very low expectations of consumer take up. For this study, we fielded separate small sample legs to assess the specific impact of a five point follow-up scale compared to a standard Dual Response scale. The results are striking.

The simple Dual Response elicited a No response on about half of all choice tasks throughout the study, with 40% of respondents indicating they would not buy the item only 4 or fewer times out of the 20 choice tasks. If we treat only "Definitely" and "Probably" accept responses to the 5 point scale question as a Yes, and the 'bottom three boxes' as a No, then a slightly higher percentage of respondents selected No in the five point scale response study leg. However, if we treat only "Definitely" as a Yes and the 'bottom four boxes' as a No, the selection of 'None' skyrockets to nearly 85%. While this level of 'None' usage may seem quite high, it is a far more accurate reflection of the true likelihood of consumers to use the new financial product. (Indeed, it is probably still too high.)

[18] In 2010, we presented at the ART Forum demonstrating a hierarchical Bayesian method integrating a choice model and a 5 point scale question (either follow-on or gathered in sequential monadic questioning prior to the choice tasks) to predict monadic ratings for concepts.

Figure 5



Proportion of 'None' selections per choice task
100%
80%
60%
40%
20%
0%
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Choice Set Sequence Number
Dual Response Y/N Model
Sequential Bottom 3 Model
Sequential Bottom 4 Model
Cumulative Proportion of 'None' selections
120%
100%
80%
60%
40%
20%
0%
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of Times NONE Selected
Dual Response Y/N Model
Sequential Bottom 3 Model
Sequential Bottom 4 Model

Final Attempt to Improve ‘None’: Combining an Expanded Scale with Dynamic Anchoring

Our goal in the subsequent studies was to attack two sources of ‘None’ selection bias independently: overstatement due to lack of an anchor and lack of discrimination in preference due to a limited scale. Our innovation was to integrate these solutions together and support this methodology with software. Specifically, our new approach involved asking respondents what brand they are currently using most often (or, alternatively, what their current behavior is), and dynamically piping in this information to a sequential response 5 point scale question at the end of every choice task.

To test the efficacy of this process, we constructed an internal study to test the potential success of a new product. As a follow-on to unrelated internal research which had identified four new high-potential product concepts in the all-natural cat litter category, we set out to test these new products against existing products. Recognizing that the category was highly fragmented, our stated goal was to contain costs while yielding the most accurate and realistic projection of likely market acceptance that we could generate.

Rather than rely on metrics of internal consistency (such as root mean squared error or out-of-sample hit rates), we rely exclusively on in-market data in order to establish empirical validity. This data was gathered from scanner data, and includes a measure of distribution (ACV%) and total unit sales volume. This data is shown below in Table 1.

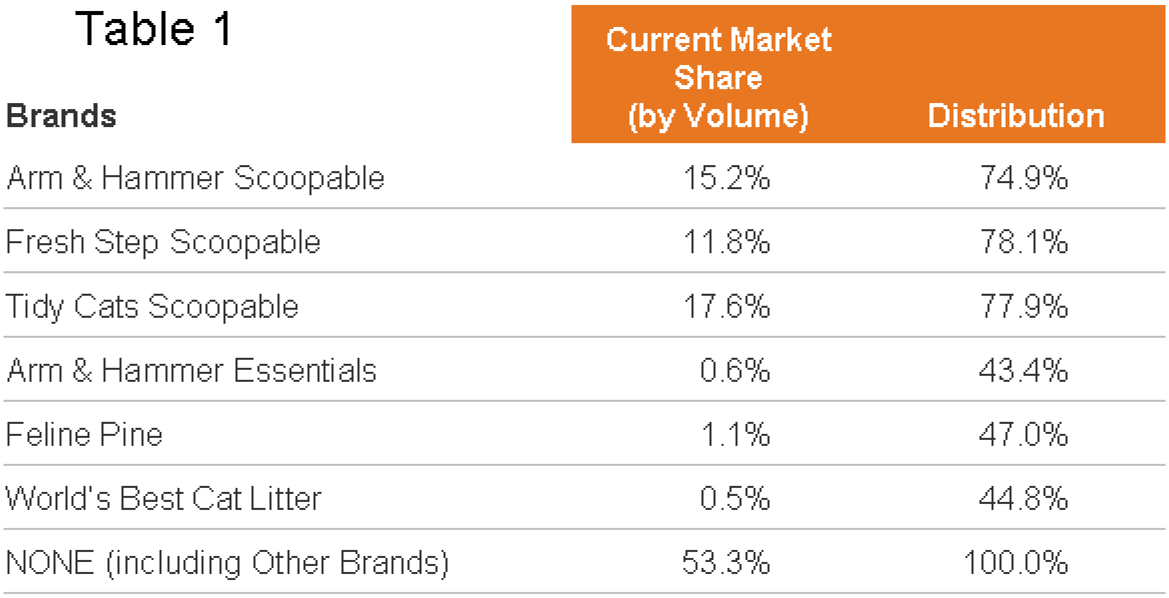
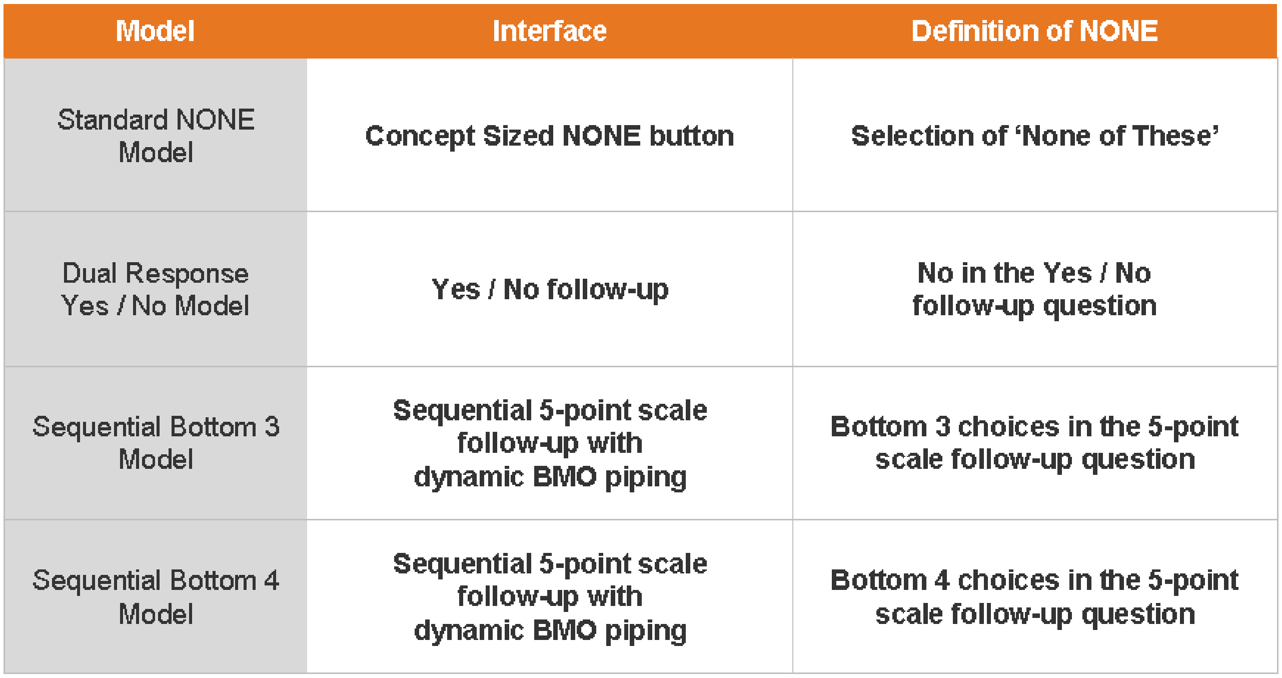


Table 1

Brands	Current Market Share (by Volume)	Distribution
Arm & Hammer Scoopable	15.2%	74.9%
Fresh Step Scoopable	11.8%	78.1%
Tidy Cats Scoopable	17.6%	77.9%
Arm & Hammer Essentials	0.6%	43.4%
Feline Pine	1.1%	47.0%
World's Best Cat Litter	0.5%	44.8%
NONE (including Other Brands)	53.3%	100.0%

In Table 2 below, we summarize the four models that were run, the three interfaces that were used (models 3 and 4 used the same interface), and the definition of a ‘None’ selection for each model. We note that in the Sequential Bottom 3 and Sequential Bottom 4 Models, when a respondent’s choice selection was the same as their current most-often-purchased product, we skipped the follow-up question and imputed a Top Box (“Definitely”) response.

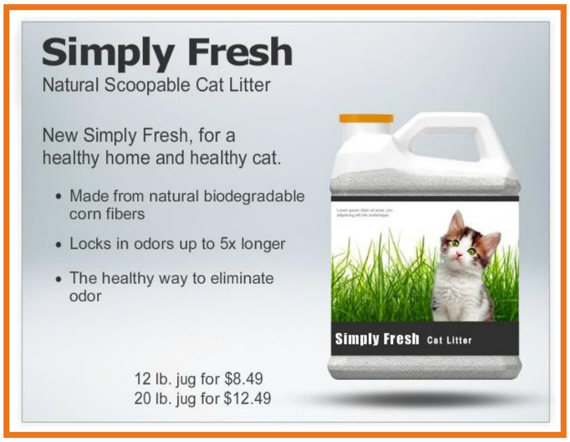
Table 2



Model	Interface	Definition of NONE
Standard NONE Model	**Concept Sized NONE button**	**Selection of ‘None of These’**
Dual Response Yes / No Model	**Yes / No follow-up**	**No in the Yes / No follow-up question**
Sequential Bottom 3 Model	**Sequential 5-point scale follow-up with dynamic BMO piping**	**Bottom 3 choices in the 5-point scale follow-up question**
Sequential Bottom 4 Model	**Sequential 5-point scale follow-up with dynamic BMO piping**	**Bottom 4 choices in the 5-point scale follow-up question**

Subsequent pages show the concept content that was tested, which include 5 test concepts and 6 current category concepts. The current category concepts were selected to provide adequate benchmarks for representing mainstream brands and the most successful all-natural brands. The design of experiment was structured to primarily show one test concept and two competitors on each page.

Test Concepts



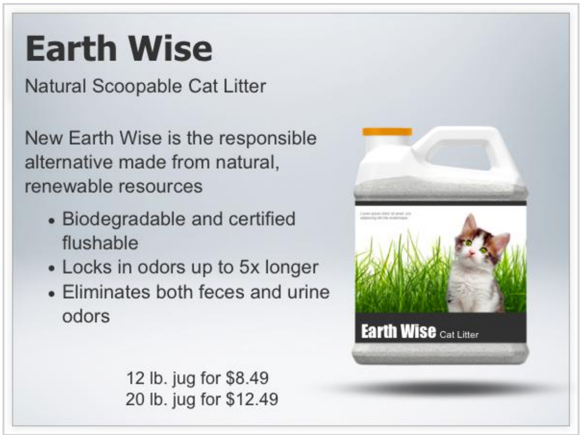
Simply Fresh

Natural Scoopable Cat Litter

New Simply Fresh, for a healthy home and healthy cat.

- Made from natural biodegradable corn fibers
- Locks in odors up to 5x longer
- The healthy way to eliminate odor

12 lb. jug for $8.49
20 lb. jug for $12.49



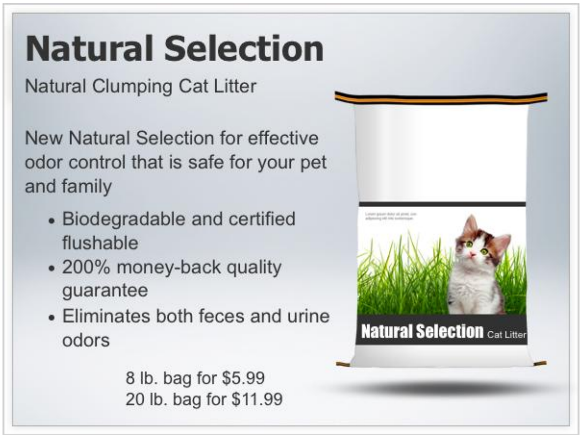
Earth Wise

Natural Scoopable Cat Litter

New Earth Wise is the responsible alternative made from natural, renewable resources

- Biodegradable and certified flushable
- Locks in odors up to 5x longer
- Eliminates both feces and urine odors

12 lb. jug for $8.49
20 lb. jug for $12.49



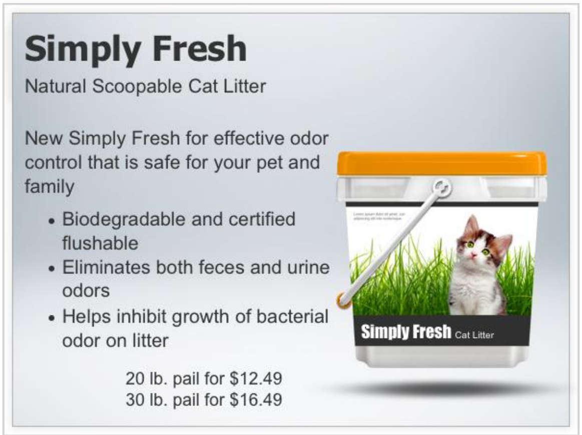
Natural Selection

Natural Clumping Cat Litter

New Natural Selection for effective odor control that is safe for your pet and family

- Biodegradable and certified flushable
- 200% money-back quality guarantee
- Eliminates both feces and urine odors

8 lb. bag for $5.99
20 lb. bag for $11.99



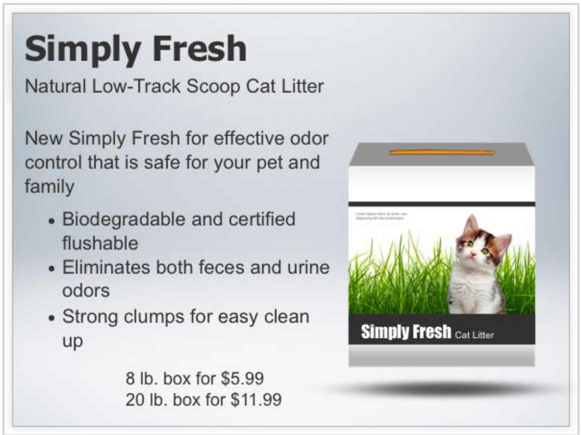
Simply Fresh

Natural Scoopable Cat Litter

New Simply Fresh for effective odor control that is safe for your pet and family

- Biodegradable and certified flushable
- Eliminates both feces and urine odors
- Helps inhibit growth of bacterial odor on litter

20 lb. pail for $12.49
30 lb. pail for $16.49



Simply Fresh

Natural Low-Track Scoop Cat Litter

New Simply Fresh for effective odor control that is safe for your pet and family

- Biodegradable and certified flushable
- Eliminates both feces and urine odors
- Strong clumps for easy clean up

8 lb. box for $5.99
20 lb. box for $11.99

Benchmark Concepts

Arm & Hammer Scoopable Cat Litter is a safe and effective way to help keep things smelling fresh and clean.

- Advanced odor control
- 99% Dust free and low tracking
- Powerful Baking Soda Crystals that eliminate odor on contact

20 lb. box for $8.89
28 lb. box for $11.20

Also available in:
- Double Duty
- Multi-Cat

Arm & Hammer Essentials harnesses the power of nature to deliver outstanding odor elimination.

- Biodegradable corn fibers
- Absorbs 2X more liquid
- Eliminates odor instantly

10.5 lb. bag for $8.30

Also available in:
- Multi-Cat

Fresh Step Premium Scoopable helps keep your home smelling fresh and your cat feeling happy.

- 99% dust free
- Anti-microbial agent helps stop odor bacteria growth
- Odor eliminating carbon for maximum odor control

20 lb. box for $9.21
25 lb. box for $10.66

Also available in:
- Multiple Cat
- Perfume & Dye Free
- Natural

Feline Pine is the smart choice for the health of your cat and your home.

- 100% natural
- 100% renewable and biodegradable
- Free of any harmful silica dust

7 lb. bag for $4.42
20 lb. bag for $8.95

Also available in:
- Clumping

Tidy Cats Scoop controls odor instantly – keeping your lively space a happy place.

- For multiple cats
- Strong clumps. Easy clean up.
- Locks in moisture – powerful odor control

14 lb. jug for $6.88
20 lb. jug for $7.58

Also available in:
- 24/7 Performance

World's Best Cat Litter, made from whole kernel corn, is 100% natural and delivers 110% performance.

- Clumps on contact
- Lasts 2 to 3 times longer than other clumping litter
- Unsurpassed odor control

8 lb. bag for $7.95

Also available in:
- Multi-Cat
- Naturally Scented

Figures 6 and 7, below, summarize the 'None' selection rates, which line up as expected. 'None' selection rates are lowest for the Standard None Model, increase substantially for the Dual Response Model, increase substantially again for the Sequential Bottom 3 Model, and increase dramatically for the Sequential Bottom 4 Model. Notably, on the Standard None Model, nearly 70% of respondents **never** selected the 'None' option, providing no information whatsoever of their threshold for participating in the category in a given shopping experience. Even with the Dual Response Model (Sequential Yes/No), over 50% of respondents **never** selected the 'None' option.

Figure 6

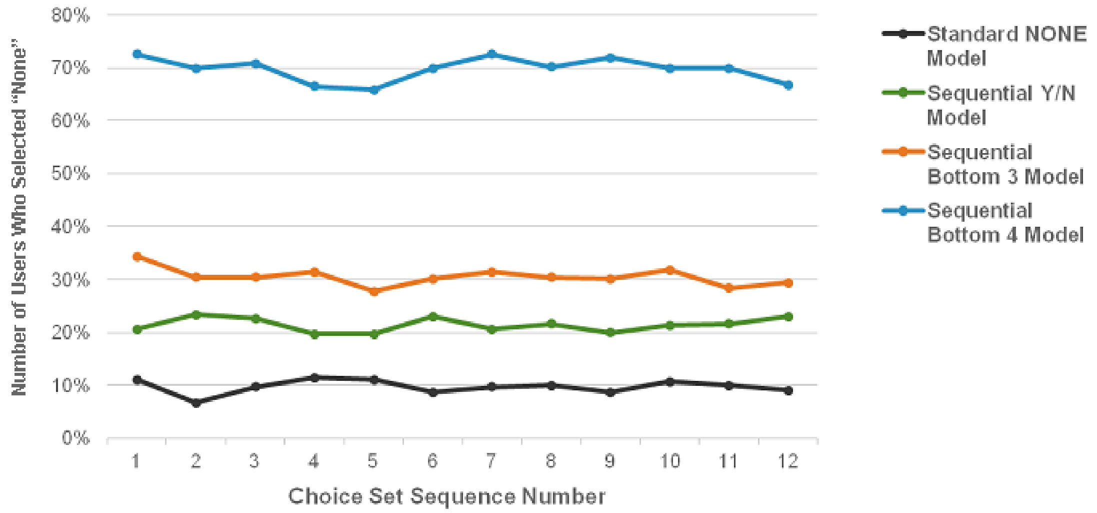
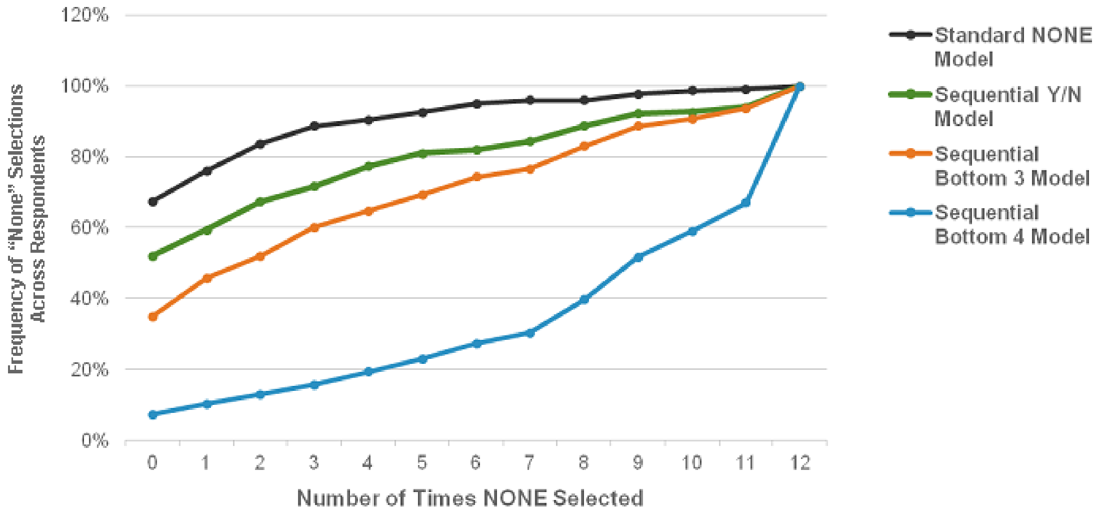


Figure 7



To demonstrate the impact of the different 'None' options, we estimated each model with a straightforward hierarchical Bayesian MNL model, and simulated preference share. In the results shown below, we directly recoded a 'None' selection as a separate parameter, and overwrote the selection of the item in the immediately prior choice task. While this eliminates relative information on inferior concepts (concepts that a consumer is not interested in buying), it offer a simple and unbiased comparison when all available concepts (including 'None') are simulated together. A more sophisticated model estimation technique is discussed later which preserves the relative information on inferior concepts.

Results are compared to actual prior year sales data. For purposes of comparing the 'None' share to an in-market benchmark, all non-included brands were summed together to reflect the fact that only current category users were included in the study. Since all included brands are being simulated, if a respondent is allocated to the 'None' option, this implies they would buy something else in the category.

Figure 8 shows the share of preference without adjustment for distribution. The most significant observation is that all of the models except for the Sequential Bottom 4 Model understate the share of 'None/Other Brands' dramatically. Likewise, all of the models except the Bottom 4 Model sharply overstate the preference share of the all natural (niche) brands.

Figure 9 shows the distribution adjusted share of preference, and leads to similar observations. The distribution adjustments modestly improve the fit of the models, however all of the models except the Bottom 4 Model continue to understate share of 'None/Other Brands' and sharply overstate the share of the niche brands. While some overstatement is expected even after distribution adjustment (smaller brands typically have worse shelf presence due to inferior shelf placement and fewer facings than larger brands), the degree of overstatement is quite large. Interestingly, applying distribution adjustments to the Bottom 4 Model causes it to *overstate* the share of 'None/Other Brands'.

Figure 8: Raw Simulated Share vs. Market Share

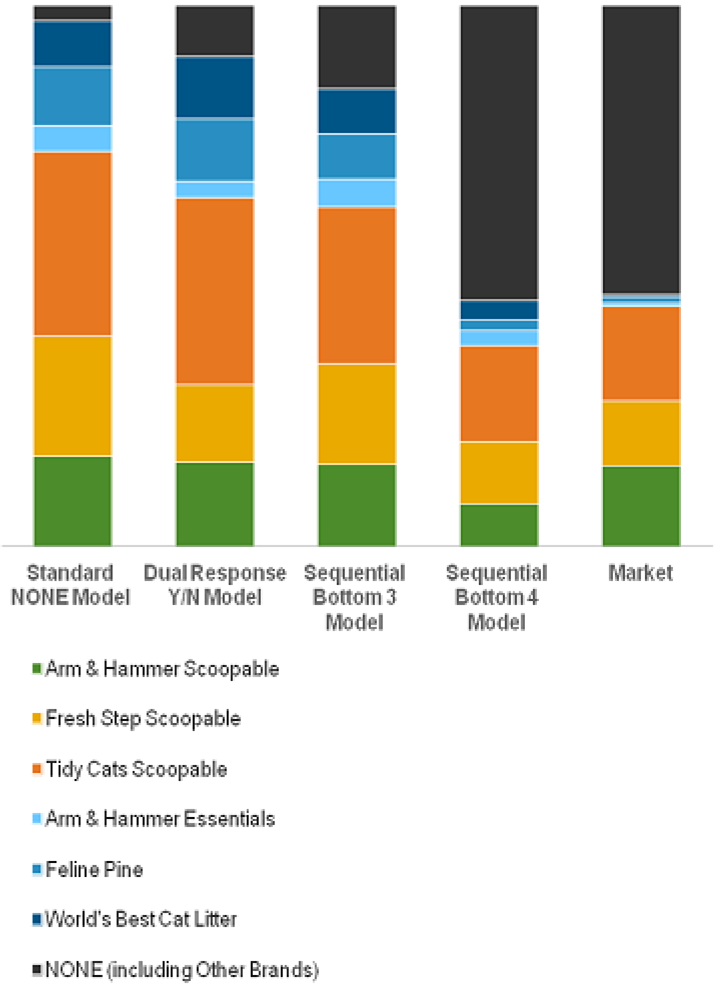
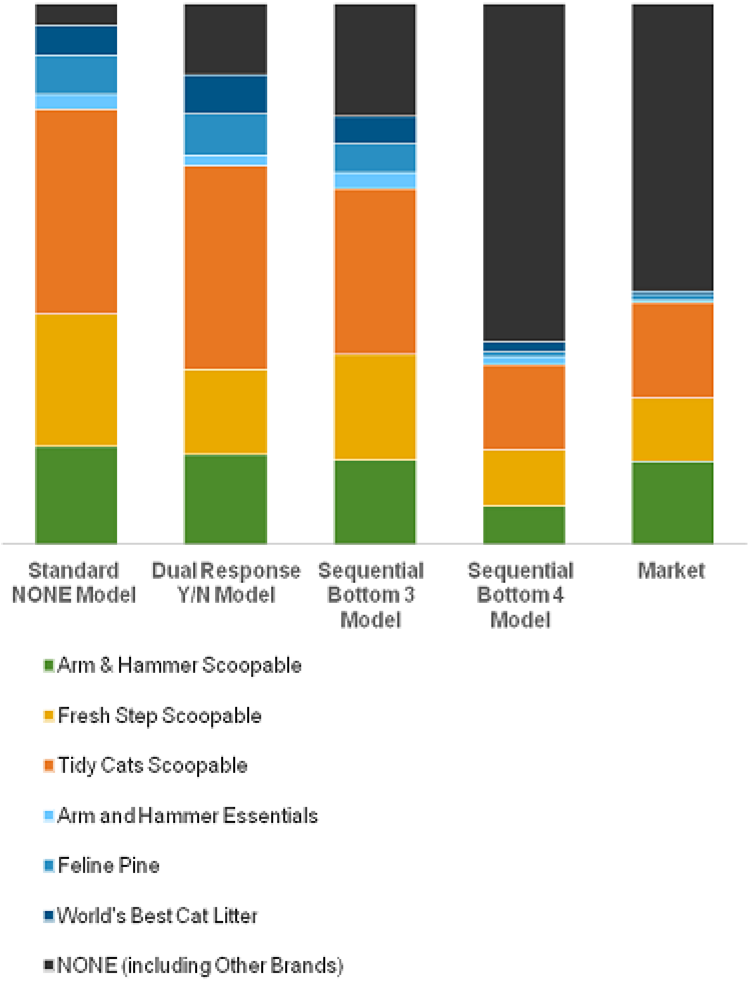


Figure 9: Distribution Adjusted Simulated Share vs. Market Share



IMPACT OF CALIBRATING MODELS THAT UNDER-PREDICT 'NONE' SHARE

One common response to models that under-predict the share of 'None/Other' is to apply calibration. A number of methods exist, all of which have drawbacks.[19] Perhaps the most frequently used approach is the application of external effects, in spite of its well known risks. We apply this method to the current data to calibrate the 'None/Other' share to the market share in order to demonstrate the impact of this technique.

One consideration in assessing the appropriateness of the use of external effects on the 'None/Other' option is the impact on individual respondent share allocation. Table 3, below, illustrates what happens when we apply external effects. The two columns separate respondents into a group which was assigned by the calibrated model to the 'None/Other' option and a group that was not assigned to the 'None/Other' option. The rows indicate which brand respondents selected as their current brand-most-often in the survey. We see considerable deviation in respondent brand usage across these groups, but the proportion of respondents who selected one of the brands that was not included in the choice study is extremely similar in both of the groups. We would generally expect to see those respondents who are assigned to 'None/Other' by the model to buy those other brands with much higher frequency in real life.

In Table 4 below, we repeat the calibration of the 'None' option to market share on the other models in our study. Although the Dual Response approach does not improve the ability of our predictive model to correlate individual 'None/Other' classification to actual purchase of other brands, both of the dynamically anchored sequential 5-point scale models do offer improvement. For the Sequential Bottom 4 Model, the improvement is quite substantial, and we observe very strong differentiation based on predicted 'None' selection. Notably, the Sequential Bottom 4 Model required the least calibration.

[19] For example, see: Orme, Bryan and Rich Johnson. "External Effect Adjustments in Conjoint Analysis". Available online in Sawtooth Software Technical Papers. Other approaches have been suggested, including post-hoc application of Bayesian constraints via a loss function: Timothy Gilbride, Peter Lenk, and Jeff Brazell (2008), "Market Share Constraints and the Loss Function in Choice Based Conjoint Analysis," Marketing Science. November/December 2008.

Table 3

Current Brand Used Most Often

	Assigned to “None” by Standard Model (with Specific Effects Adjustments)?	
	NOT Assigned to “None”	Assigned to “None”
Arm & Hammer Scoopable	8%	14%
Fresh Step Scoopable	20%	20%
Tidy Cats Scoopable	25%	11%
Arm & Hammer Essentials	3%	2%
Feline Pine	3%	1%
World's Best Cat Litter	0%	3%
Other Brands (“None”)	**42%**	**49%**
Sample Size	154	146

Table 4

Current Brand Used Most Often

	Standard NONE Model		Dual Response Y / N Model		Sequential Bottom 3 Model		Sequential Bottom 4 Model	
Most Often Brand	A	B	A	B	A	B	A	B
Arm & Hammer Scoopable	8%	14%	13%	10%	12%	11%	14%	9%
Fresh Step Scoopable	20%	20%	11%	14%	26%	17%	28%	15%
Tidy Cats Scoopable	25%	11%	21%	20%	21%	22%	28%	15%
Arm & Hammer Essentials	3%	2%	9%	3%	4%	5%	7%	1%
Feline Pine	3%	1%	0%	1%	1%	0%	0%	1%
World's Best Cat Litter	0%	3%	1%	4%	1%	0%	1%	1%
Other Brands (“None”)	**42%**	**49%**	**45%**	**49%**	**35%**	**46%**	**23%**	**58%**
Sample Size	154	146	159	141	144	156	149	151

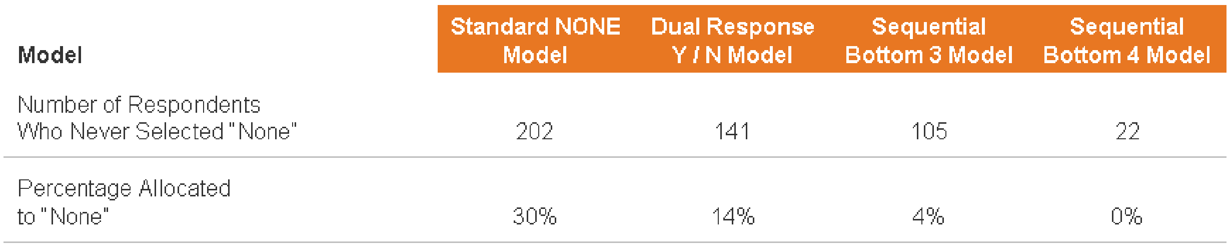
Segment A: Respondents for whom the highest utility is on one of the products included in the study

Segment B: Respondents for whom the highest utility is on "None"

Table 5 below further illustrates the risks of applying market calibration to adjust for the overstatement of purchase intent caused by under-use of the ‘None’ option. For each model, we tabulate the number of respondents who never selected ‘None’ who are predicted to select ‘None’ by our calibrated model. All of the models except the Sequential Bottom 4 Model assign at least a modest percentage of respondents to ‘None’ even though they never selected ‘None’ in their choice experience.

If we recall Figure 6, we note that nearly 70% of respondents who participated in the Standard None Model never actually selected ‘None’ in any of their choice tasks. Thus, our calibrated model standard model is assigning many people to ‘None’ without any basis whatsoever except using an imputed threshold for ‘None’ derived via hierarchical Bayesian estimation. Indeed, in order to limit the number of individuals who are assigned to ‘None’ who never selected it to only 30% of the total, the Standard None Model is also assigning virtually everyone who ever selected ‘None’ in their choice exercise to the ‘None’ option.

Table 5

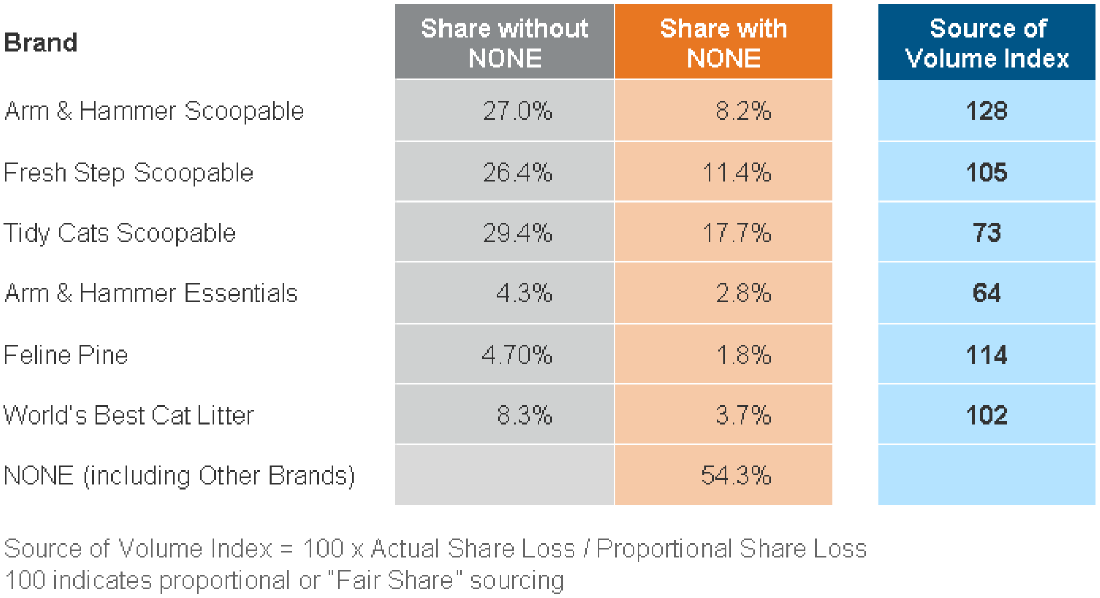


Model	Standard NONE Model	Dual Response Y / N Model	Sequential Bottom 3 Model	Sequential Bottom 4 Model
Number of Respondents Who Never Selected "None"	202	141	105	22
Percentage Allocated to "None"	30%	14%	4%	0%

Finally, in assessing the impact of application of external effects adjustment to calibrate the data to market, we also consider the degree to which ‘None’ is sourcing volume non-proportionately from other brands. Even if individual level assignment was extremely poor, it’s possible for aggregate level projections to be accurate. The risks of improper aggregate level findings would decrease if the ‘None’ option in our model was sourcing proportionately from other brands. As Table 6 below shows, however, the ‘None’ option is not sourcing proportionately.

The first column shows the projected share without the inclusion of the ‘None’ option. The second column shows the share with the inclusion of the ‘None’ option, and the final column calculates a source of volume index. This index is constructed such that 100 indicates the new brand (‘None’) is pulling share from existing brands in proportion to their current share. As we would expect, it is not. Indeed, ‘None’ is undersourcing from brands such as Tidy Cats Scoopable and Arm & Hammer Essentials, and oversourcing from Arm & Hammer Scoopable.

Table 6



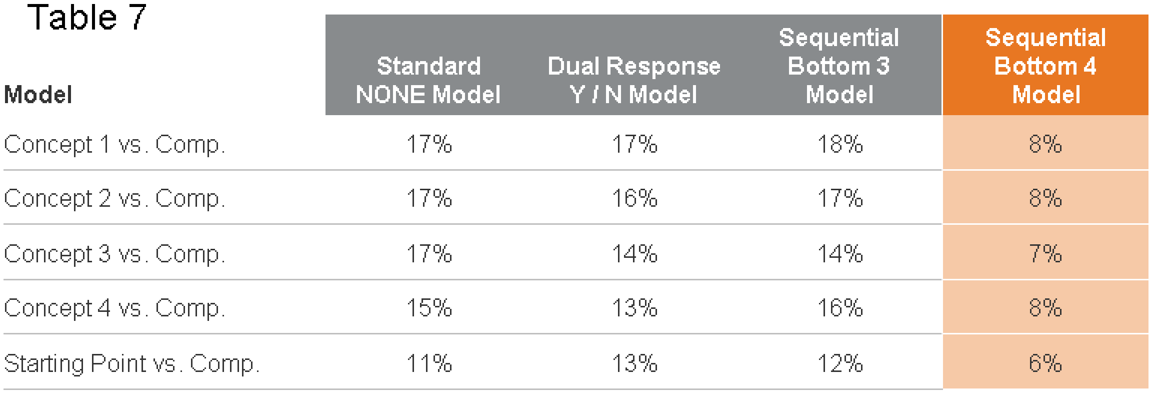
Brand	Share without NONE	Share with NONE	Source of Volume Index
Arm & Hammer Scoopable	27.0%	8.2%	**128**
Fresh Step Scoopable	26.4%	11.4%	**105**
Tidy Cats Scoopable	29.4%	17.7%	**73**
Arm & Hammer Essentials	4.3%	2.8%	**64**
Feline Pine	4.70%	1.8%	**114**
World's Best Cat Litter	8.3%	3.7%	**102**
NONE (including Other Brands)		54.3%	

Source of Volume Index = 100 x Actual Share Loss / Proportional Share Loss
100 indicates proportional or "Fair Share" sourcing

Impact of Model Selection on Test Concept Scores

One critical question in assessing the models asks about the impact of the 'None' approach on our projections for the new test products we included in the study (the putative objective of our internal client). To answer this question, we simulate the introduction of each of the test concepts vs. all competitors (and the 'None') option. As shown below in Table 7, all of the first three models yield substantially higher share projections than the Sequential Bottom 4 Model. In addition, the relative ordering of the test concepts differ. Concept 4, for example, underperforms in the Standard None Model, but outperforms in the Sequential Bottom 4 Model. All of the models, however, do detect a lift vs. the Starting Point concept, which was expected.[20]

Table 7



Model	Standard NONE Model	Dual Response Y / N Model	Sequential Bottom 3 Model	Sequential Bottom 4 Model
Concept 1 vs. Comp.	17%	17%	18%	8%
Concept 2 vs. Comp.	17%	16%	17%	8%
Concept 3 vs. Comp.	17%	14%	14%	7%
Concept 4 vs. Comp.	15%	13%	16%	8%
Starting Point vs. Comp.	11%	13%	12%	6%

[20] The Starting Point concept reflected the best guess of our internal marketing team prior to undertaking a concept optimization exercise. Part of the objective of this study was to assess the lift generated by the optimization process. This was conducted separately from the research discussed in this paper.

FURTHER IMPROVEMENTS

One challenge to some of the models above, and particularly to the Sequential Bottom 4 Model, is the information loss by simply recoding 'None' selections by overwriting the selection of the concept selected prior to the Dual Response Yes/No or Sequential Response 5-point Scale. This was discussed by Orme (see above) and addressed by recoding the Dual Response dataset in such a manner as to fully utilize the relative choice data and the dual response data.

In Figure 10, we outline an alternate method of addressing this concern for the Sequential Bottom 4 Model. This method was discussed in our 2010 ART Forum Presentation.[21] In essence, we posit a data generating model (DGM) in which the choice outcome and the scale rating outcome are both stochastic manifestations of a unified underlying utility structure. The choice data reflects assumptions similar to those in a multinomial logistic model (for simplicity), and the scale data reflects a threshold model similar to an ordered logistic model. With 5 possible ordered outcomes, there are 4 cutoff points for each individual respondent. Cutoff points can be estimated for all respondents individually and subjected to a distribution, however this is empirically difficult and unstable. This paper uses a simpler and more stable method in which cutoffs are fitted at the aggregate level, but each respondent is permitted a single adjustment factor which is estimated by augmenting the Beta draw matrix in a hierarchical Bayesian model. We note that it is often reasonable to collapse categories, such as the bottom three boxes of a model, since predicting rank selection beyond the top box (Definitely) and the second box (Probably) is empirically irrelevant.

Once the model described in Figure 10 is estimated, it can be used to predict both share allocation and likelihood of selecting a Top Box (Definitely) or Second Box (Probably) rating for that item. With this data, it is a simple exercise to apply empirical weighting; for example, we may assume that someone who indicates they would definitely buy the product instead of the product they currently buy has a true purchase probability of 100%, and that someone who indicates they would probably buy the product instead of the product they current buy most often has a purchase probability of 20%. The selection of these weighting factors is entirely normative.

Figure 11 shows the empirical results of using the weighting method described above. Given that we are already simulating all of the available concepts in the test, the inclusion of all of the choice data in the model does not affect the outcome that much (second-best data indicating what each buyer would do if one of the brands was not available is less relevant, except for improving the covariance matrix estimation in the HB model). The primary benefit of weighting the data based on Top or Second Box prediction is that we effectively create a weighted average of the Sequential Bottom 3 Model and the Sequential Bottom 4 Model. The final distribution adjusted model comes remarkably close to predicting in-market share, without any further arbitrary calibration.

[21] Karty, Kevin. "Fusing Choice and Rating Scale Data to Predict Rating Outcomes" ART Forum 2010.

Figure 10

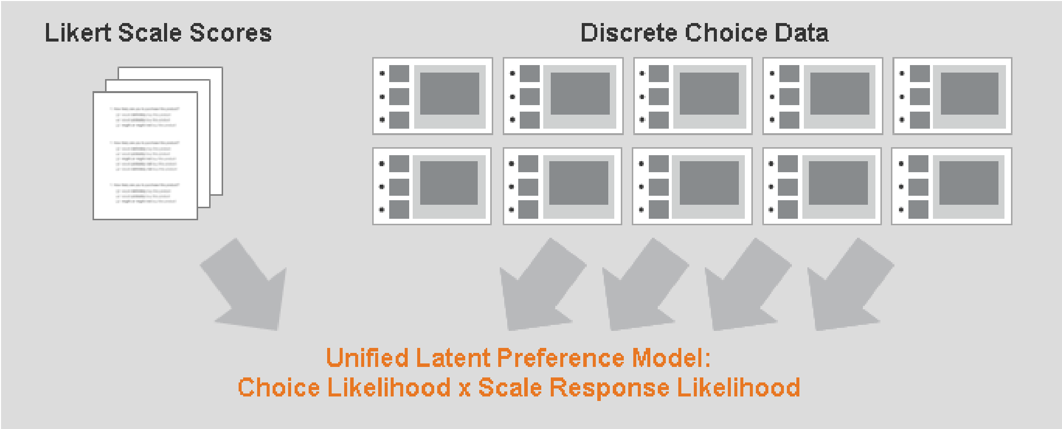
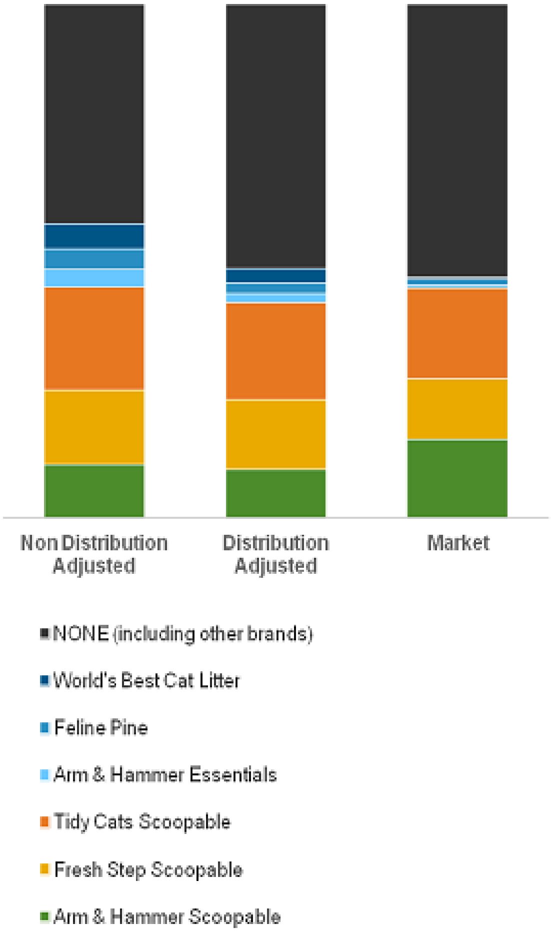

Likert Scale Scores
Discrete Choice Data
Unified Latent Preference Model:
Choice Likelihood x Scale Response Likelihood

Figure 11: Distribution Adjusted Simulated Share vs. Market Share



CONCLUSION AND DISCUSSION

We believe we have demonstrated an approach to dramatically improve the empirical reliability and predictive power of choice models that rely heavily on the 'None' option. It assess two known problems with the 'None' option – non-specificity of the 'None' resulting from the lack of an anchor, and excessive scale compression. Our approach uses a dynamically generated anchor provided by the respondent earlier in the survey, and expands the sequential response scale from the 2-point scale used in a standard Dual Response model to a 5-point scale. Both improvements seem to contribute to the improved outcome, though we do not quantitatively measure the importance of each separate modification.

Although initial outcomes appear compelling, we note several remaining weaknesses in the approach. First, our primary test category was consumer package goods, and we do not know how well this method expands to other categories or sectors – such as consumer electronics, credit cards, or pharmaceuticals. Second, we have not thoroughly vetted this process for use in main effects models with many attributes. Our primary applications remain in consumer package goods, although we have observed apparent success throughout our business domains. Third, obtaining full benefit from this approach may require the use of normative weights on sequential response rating outcomes, much like in monadic testing, and it is well known that these weights can vary significantly across categories, cultures, and countries. Clearly, significant work remains to fully assess the potential of this approach. Nonetheless, the method also appears to offer a near term opportunity to improve the empirical validity and quality of our modeling projections.

The Voice of the Patient [22]

Christopher Saigal
Ely Dahan
UCLA

Abstract:

We present a voice-of-the-customer-based method for identifying conjoint attributes and levels directly from customer interviews. This approach elicits non-obvious attributes, gives voice to customer concerns, and applies Agglomerative Hierarchical Clustering (AHC) to quantify attributes and levels that easily plug into conjoint. This approach gathers and organizes the voice of the patient for the purpose of attribute-based preference measurement of alternative healthcare treatments. The method borrows heavily from the voice of the customer literature in marketing, with appropriate adaptations to the medical context. A real-world application highlights the method and its benefits, and initial results show promise in terms of eliciting non-obvious attributes, fairly characterizing key patient concerns, and helping to generate a useful list of attributes and levels.

The Voice of the Patient:

When patient preferences for treatment alternatives need to be measured, health care researchers must first develop a compact list of treatment attributes and their possible levels for patients to consider. The quality and comprehensiveness of such an attribute list is crucial to the success of preference measurement methods such as conjoint analysis, ratings or time-tradeoff.

Traditionally in medicine, such lists of attributes and levels are developed based on the expertise of the medical researchers and the clinicians who provide treatment. The analogy in product marketing, where preference measurement is used extensively, would be to have marketing managers develop the list of product or service attributes, and the levels of each attribute, based on prior experience.

Griffin and Hauser (1993), in their seminal work on the "The Voice of the Customer," highlight the value of having attributes and levels emerge organically, directly from the voices of customers rather than managers. They lay out a method for objectively gathering and organizing customer statements about products and services. The present research seeks to apply their approach to health care, and to adapt it to the special constraints and requirements found in medical research.

[22] This report describes research done within UCLA's Urology Research Group within the Department of Urology at UCLA's Geffen School of Medicine. This research was sponsored by a grant from the National Institute of Health under contract No. R01CA134997-01A1. Comments on the research from Dr. Robert Kaplan, Dr. Mark Litwin and Dr. Kate Crespi have been extremely helpful as have the comments and suggestions from 16 "Voice of the Patient" interviewees, 18 sorters, participants in three focus groups and 17 initial subjects who completes the survey instruments, and the research assistance of Sylvia Lambrechts and Elizabeth Garcia. Authors emails: elydahan@gmail.com, csaigal@mednet.ucla.edu.

PRIOR RESEARCH:

Collecting and organizing the voice of the customer has been an integral part of the method of Quality Function Deployment (QFD). A key element of the approach is to maintain the customer voice in unfiltered, verbatim form so that attributes can be clearly understood and avoid researcher biases. Once the customer voice has been parsed into individual statements, each pertaining to a particular need, these can be organized based on similarity using perceived affinity. We propose to adapt the elements of verbatim data and affinity sorting to the health care research context by showing representative quotations for each attribute to all users of the preference measurement methods being developed.

Burchill and Hepner-Brodie (2005) propose methods of translating the voice of the customer into actions to improve purchase probability and customer satisfaction. The present research contributes to the existing literature by (1) quantitatively clustering the sorted verbatim statements using Agglomerative Hierarchical Clustering and, (2) adapting the methods to the medical context.

Differences in Medicine:

Preference measurement in healthcare differs from that for product marketing in important ways. First, new product marketers are primarily focused on the attributes and levels that potential consumers want or need, while healthcare research may focus more on treatment factors and outcomes that patients most want to avoid. This is driven by the fact that product purchases are typically optional, while healthcare decisions are, in most cases, required. Given that perfect health is the basis of comparison in most healthcare preference measurement scenarios, all other states are necessarily somewhat less desirable.

This leads to treatment attribute levels that range from the perfect health norm to lesser levels that represent various lower levels of health, symptoms, or treatment consequences. Thus, while voice of the customer interviews might cover the attributes and levels that make potential customers more likely to buy, voice of the patient interviews might instead focus on treatment features and outcomes that are most troublesome or to be avoided altogether.

Another key difference is that while in marketing customers and marketing managers may generally agree on the key attributes for any given product or service (although not necessarily on which are most important), in medicine the priorities of providers may vary dramatically from those of patients because the two groups view treatments from unique perspectives.

Professionals focus on patient health, treatment outcomes, and quality of care, while patients may emphasize emotional response, convenience, and lifestyle effects. Of course, in many cases these objectives align, but possibly less than perfectly. This argues not only for interviewing patients to gather their verbatims, but also for having fellow patients sort these quotations by affinity.

METHODOLOGY OVERVIEW:

We proceed to gather, parse, cull, organize and cluster the voice of the patients along the seven key steps depicted in Figure 1.

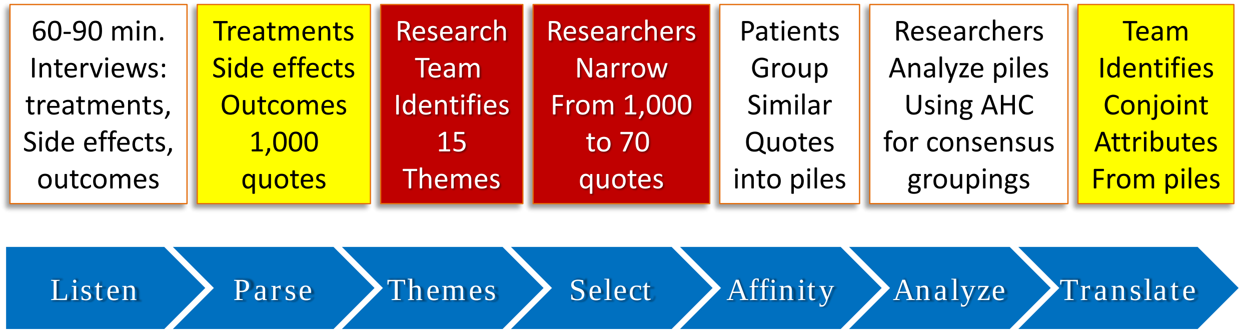

60-90 min. Interviews: treatments, Side effects, outcomes
Treatments Side effects Outcomes 1,000 quotes
Research Team Identifies 15 Themes
Researchers Narrow From 1,000 to 70 quotes
Patients Group Similar Quotes into piles
Researchers Analyze piles Using AHC for consensus groupings
Team Identifies Conjoint Attributes From piles
Listen
Parse
Themes
Select
Affinity
Analyze
Translate

Figure 1: Translating the Voice-of-the-Customer ➔ Conjoint Attributes

The three steps depicted in **black & white** in Figure 1 (steps 1, 5 and 6) are largely objective and require minimal researcher judgment, while the two steps in **yellow** (steps 2 and 7) require a *small* degree of researcher judgment and the two **red steps** (steps 3 and 4) a *significant* amount of subjective researcher judgment. In order to reduce errors in the more subjective steps, we recommend (*and utilized*) independent judgments by multiple judges, and consensus amongst those judges after making their independent judgments. The reasoning and methodology behind each step is described below.

Step 1: Listen through Individual Interviews or Web-based Text Mining

At least 15-20 individual interviews are conducted and recorded word-for-word, or internet-based text is mined. Researchers should try not to bias the results, but can still keep the focus on potential needs, wants, and concerns. A thorough interview might last 30-90 minutes. Videotaped interviews can also work as long as they are in a digitized format that allows for segments to be separated. We conducted 17 such interviews with individual prostate cancer patients.

Step 2: Parse the Interviews into Verbatims

The interview transcripts (or video clips) are then parsed into individual quotations, each of which focuses on a particular product or service need, want or concern. 20 customer interviews could easily generate 500 to 1,000 individual quotations. A spreadsheet or database program with sorting capability provides the ideal storage medium for the parsed quotations as they will be reordered and reorganized frequently. In our prostate cancer treatment research, the 17 individual interviews produced 907 parsed quotations, 474 of which were deemed to relate to prostate cancer treatment.

Step 3: Develop Possible Themes from the Verbatims

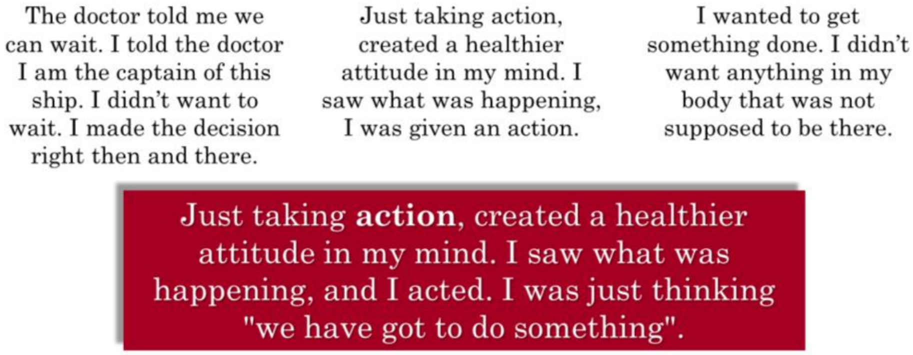
Multiple researchers independently evaluate the full set of quotations in an attempt to identify a superset of potential affinity-based "themes." Researchers then achieve consensus on a set of possible themes numbering approximately two to three times the number of attributes that will ultimately be identified. Themes capture the underlying element that a particular set of quotations have in common, even if the quotations hold varying points of view on that topic.

Based on 474 quotations, three independent judges developed the following (15) fifteen themes:

1. Back-up Plan
2. Being cautious
3. Bowel problems
4. Don't cut me
5. Get it out of me
6. Hurting non-prostate areas; Precision Targeting
7. Living longer
8. Other men's prostate cancer experience
9. Others Influence (Doctors, Family, Friends)
10. Sex life
11. Staying independent; don't want to be a burden
12. Taking action
13. Urinary Incontinence
14. Will I be worried all the time
15. Will it come back

Step 4: Select Representative Verbatim Quotations

Utilizing the themes developed in step 3, multiple researchers independently identify a smaller subset of the 474 quotations from step 2 that *best capture* each theme. The researchers then achieve consensus on 3-7 quotations that most fully capture the essence of each theme, and also balance numerous customer points-if-view. In some cases, as shown below in Figure 2, several quotations can be combined and condensed to capture the essential elements from the voice of the patient.



The doctor told me we can wait. I told the doctor I am the captain of this ship. I didn't want to wait. I made the decision right then and there.

Just taking action, created a healthier attitude in my mind. I saw what was happening, I was given an action.

I wanted to get something done. I didn't want anything in my body that was not supposed to be there.

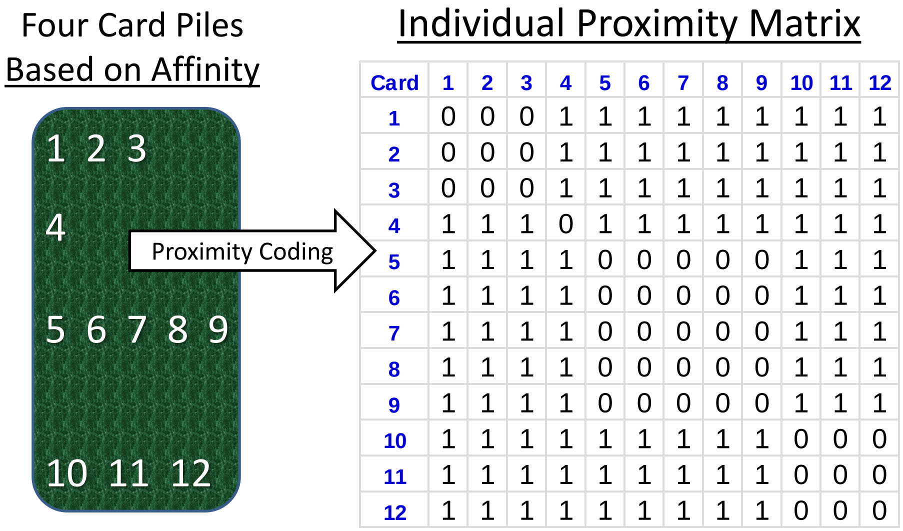
Just taking **action**, created a healthier attitude in my mind. I saw what was happening, and I acted. I was just thinking "we have got to do something".

Figure 2: Narrowing from 474 Quotations to the 70 *Most Representative* Ones

This subset of representative quotations form a compact "deck" of quotations cards which can now be sorted by a new set of potential customers based on affinity between the quotations in the next step. In our case, we narrowed to a 70-quotation deck, or approximately 15% of the 474-element superset of treatment-related quotations.

Step 5: Fellow Customers Express Affinity between Verbatims

A new group of 20-100 subjects is invited to individually place quotations from the sorting deck into piles based on perceived affinity. If the task is too challenging to perform with no prompting, as it proved to be in our case, title cards representing each theme developed in step 3 can be placed on the sorting table to facilitate sorting based on affinity. The piles created by each individual can be coded into a proximity matrix as shown in Figure 3 below.

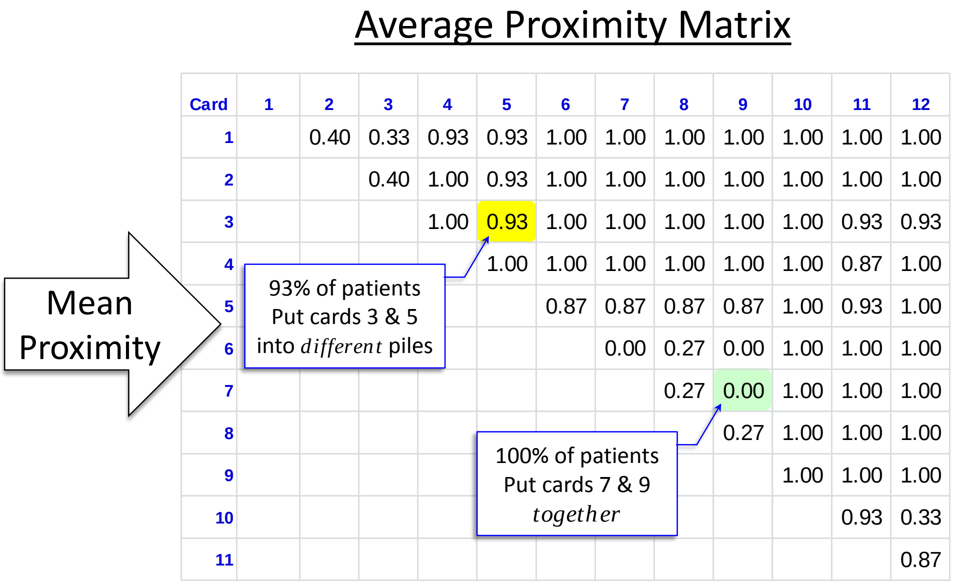


Card	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	1	1	1	1	1	1	1	1	1
2	0	0	0	1	1	1	1	1	1	1	1	1
3	0	0	0	1	1	1	1	1	1	1	1	1
4	1	1	1	0	1	1	1	1	1	1	1	1
5	1	1	1	1	0	0	0	0	0	1	1	1
6	1	1	1	1	0	0	0	0	0	1	1	1
7	1	1	1	1	0	0	0	0	0	1	1	1
8	1	1	1	1	0	0	0	0	0	1	1	1
9	1	1	1	1	0	0	0	0	0	1	1	1
10	1	1	1	1	1	1	1	1	1	0	0	0
11	1	1	1	1	1	1	1	1	1	0	0	0
12	1	1	1	1	1	1	1	1	1	0	0	0

Figure 3: Sample Proximity Matrix and Distance Coding for an Individual grouping 12 quotations into four piles based on affinity

Step 6: Agglomerative Hierarchical Clustering

The individual proximity matrices are averaged across all respondents. The mean distances represent the average distance between any two quotations, and are also a measure of the probability that two quotations are in the same pile or in different piles. For example, a 0.2 mean proximity means that 80% of respondents put that pair of quotations in the same theme pile while 20% put the quotations into different themes. In other words, these distances are objective measures of affinity.


Average Proximity Matrix
Mean
Proximity

Card	1	2	3	4	5	6	7	8	9	10	11	12
1		0.40	0.33	0.93	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2			0.40	1.00	0.93	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3				1.00	0.93	1.00	1.00	1.00	1.00	1.00	0.93	0.93
4					1.00	1.00	1.00	1.00	1.00	1.00	0.87	1.00
5						0.87	0.87	0.87	0.87	1.00	0.93	1.00
6							0.00	0.27	0.00	1.00	1.00	1.00
7								0.27	0.00	1.00	1.00	1.00
8									0.27	1.00	1.00	1.00
9										1.00	1.00	1.00
10											0.93	0.33
11												0.87

93% of patients Put cards 3 & 5 into *different* piles

100% of patients Put cards 7 & 9 *together*

Figure 4: The Mean Proximity Matrix across all Subjects for the first 12 quotations

The average proximity matrix is then analyzed using Agglomerative Hierarchical Clustering (AHC) to group quotations into a hierarchical tree, a schematic of which appears in below in Figure 5.

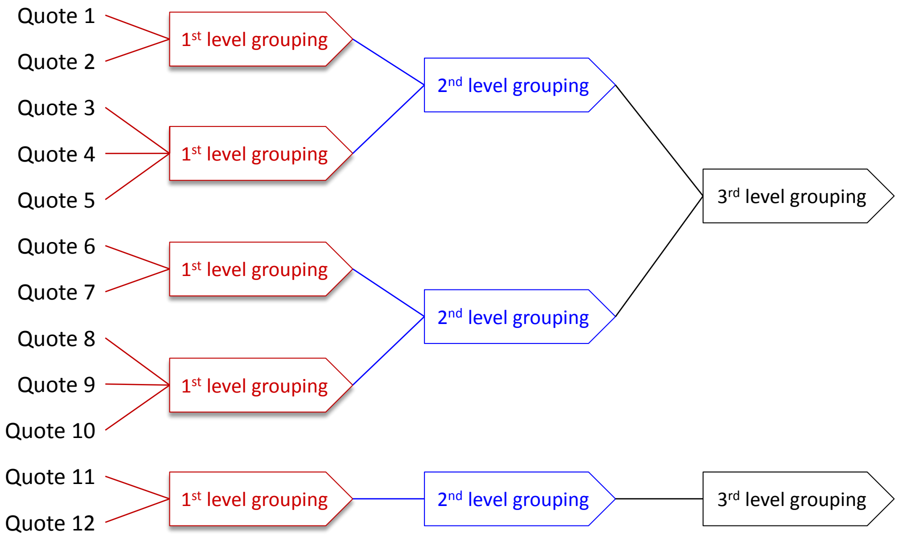
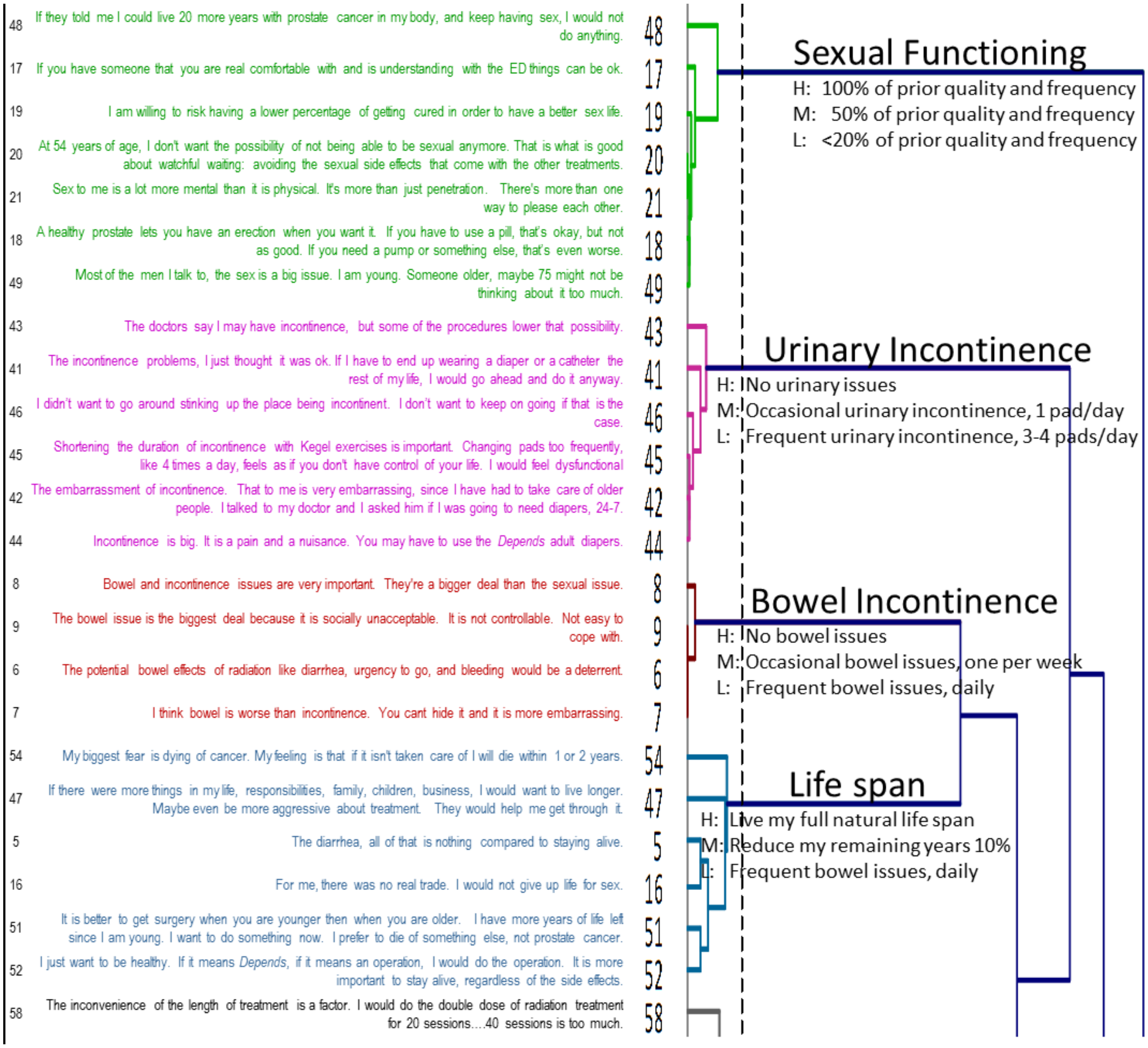


Figure 5: Agglomerative Hierarchical Clustering (AHC) Tree Structure

The first level groupings connect quotations that are perceived to have high affinity, while higher level groupings aggregate these smaller clusters based on their affinity with each other. The final number of groupings to be used in defining conjoint attributes depends on pragmatic issues such as how many attributes conjoint subjects can handle cognitively and how many part worths the conjoint methodology can reasonably estimate at the individual level. Figure 6 illustrates a portion of the AHC tree generated for the 70 prostate cancer quotations, and shows where the tree was cut off when subjectively identifying four of the seven conjoint attributes.



48 If they told me I could live 20 more years with prostate cancer in my body, and keep having sex, I would not do anything.

17 If you have someone that you are real comfortable with and is understanding with the ED things can be ok.

19 I am willing to risk having a lower percentage of getting cured in order to have a better sex life.

20 At 54 years of age, I don't want the possibility of not being able to be sexual anymore. That is what is good about watchful waiting: avoiding the sexual side effects that come with the other treatments.

21 Sex to me is a lot more mental than it is physical. It's more than just penetration. There's more than one way to please each other.

18 A healthy prostate lets you have an erection when you want it. If you have to use a pill, that's okay, but not as good. If you need a pump or something else, that's even worse.

49 Most of the men I talk to, the sex is a big issue. I am young. Someone older, maybe 75 might not be thinking about it too much.

43 The doctors say I may have incontinence, but some of the procedures lower that possibility.

41 The incontinence problems, I just thought it was ok. If I have to end up wearing a diaper or a catheter the rest of my life, I would go ahead and do it anyway.

46 I didn't want to go around stinking up the place being incontinent. I don't want to keep on going if that is the case.

45 Shortening the duration of incontinence with Kegel exercises is important. Changing pads too frequently, like 4 times a day, feels as if you don't have control of your life. I would feel dysfunctional

42 The embarrassment of incontinence. That to me is very embarrassing, since I have had to take care of older people. I talked to my doctor and I asked him if I was going to need diapers, 24-7.

44 Incontinence is big. It is a pain and a nuisance. You may have to use the *Depends* adult diapers.

8 Bowel and incontinence issues are very important. They're a bigger deal than the sexual issue.

9 The bowel issue is the biggest deal because it is socially unacceptable. It is not controllable. Not easy to cope with.

6 The potential bowel effects of radiation like diarrhea, urgency to go, and bleeding would be a deterrent.

7 I think bowel is worse than incontinence. You cant hide it and it is more embarrassing.

54 My biggest fear is dying of cancer. My feeling is that if it isn't taken care of I will die within 1 or 2 years.

47 If there were more things in my life, responsibilities, family, children, business, I would want to live longer. Maybe even be more aggressive about treatment. They would help me get through it.

5 The diarrhea, all of that is nothing compared to staying alive.

16 For me, there was no real trade. I would not give up life for sex.

51 It is better to get surgery when you are younger then when you are older. I have more years of life left since I am young. I want to do something now. I prefer to die of something else, not prostate cancer.

52 I just want to be healthy. If it means *Depends*, if it means an operation, I would do the operation. It is more important to stay alive, regardless of the side effects.

58 The inconvenience of the length of treatment is a factor. I would do the double dose of radiation treatment for 20 sessions....40 sessions is too much.

48
17
19
20
21
18
49
43
41
46
45
42
44
8
9
6
7
54
47
5
16
51
52
58
Sexual Functioning
H: 100% of prior quality and frequency
M: 50% of prior quality and frequency
L: <20% of prior quality and frequency
Urinary Incontinence
H: No urinary issues
M: Occasional urinary incontinence, 1 pad/day
L: Frequent urinary incontinence, 3-4 pads/day
Bowel Incontinence
H: No bowel issues
M: Occasional bowel issues, one per week
L: Frequent bowel issues, daily
Life span
H: Live my full natural life span
M: Reduce my remaining years 10%
L: Frequent bowel issues, daily

Figure 6: Quotations Organized by AHC into a Hierarchical Tree

Step 7: Researchers Translate Clusters into Attributes and Levels

The researchers translate Figure 6's tree structure of quotations into conjoint attributes, as shown in Figure 7, and then into attribute levels as in Figure 8. By directly connecting the attributes and levels to this objectively organized version of patients' voices, the probability that conjoint subjects will "connect" to the study improves.

Treatment Issues	Side Effects
Cutting: I don't want to be cut; I don't want to have surgery.	*Sex*: If you have an understanding partner, the ED thing can be ok.
Others' Advice: I only follow doctors' advice up to a point. Not 100%	*Urinary*: Changing pads frequently…feels as if you don't have control of your life.
Caution: I could wait for a while if the numbers stay stable…	*Lifespan*: It is more important to stay alive, regardless of the side effects.
Action: I was just thinking "we have got to do something"	*Bowel*: The bowel issue is the biggest deal because it is socially unacceptable.

Figure 7: Interpreting the AHC Output as Attributes and representing each attribute with best-in-cluster quotations

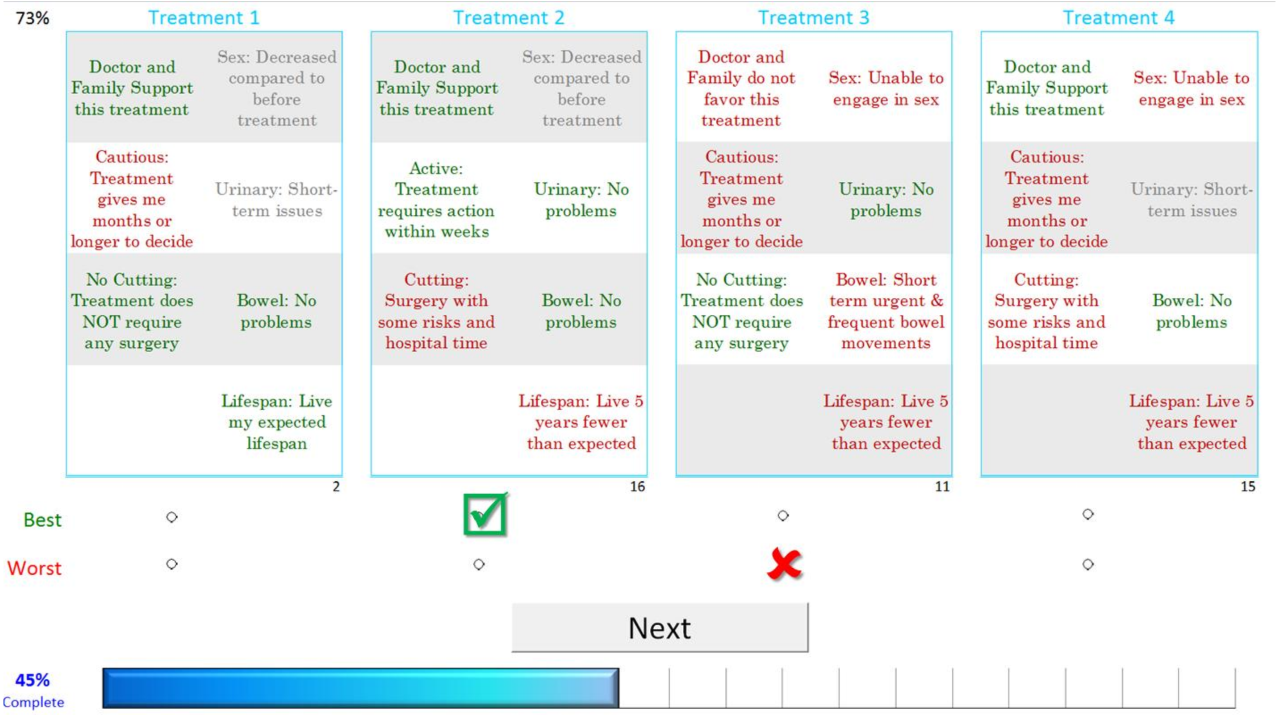
The final list of conjoint attributes and levels is based on the tree from step 6, but also need to be worded to be easily and quickly understood by conjoint subjects.

	Attribute Level1	*Attribute Level2*	*Attribute Level3*
Others Support	*Doctor and Family Support this treatment*	*Doctor and Family do not favor this treatment*	
Action/ Caution	*Active: Treatment requires action within weeks*	*Cautious: Treatment gives me months or longer to decide*	
Surgery	*No Cutting: Treatment does NOT require any surgery*	*Cutting: Surgery with some risks and hospital time*	
Sex	*Sex: Same as before treatment*	*Sex: Decreased compared to before treatment*	*Sex: Unable to engage in sex*
Urinary	*Urinary: No problems*	*Urinary: Short-term issues*	*Urinary: Long-term issues*
Bowel	*Bowel: No problems*	*Bowel: Short term urgent & frequent bowel movements*	
Lifespan	*Lifespan: Live my expected lifespan*	*Lifespan: Live 5 years fewer than expected*	

Figure 8: Seven Attributes with 2 or 3 levels each

The factors affecting the decision about how many attributes, and the number of levels within each attribute, are largely outside the scope of the present research, but include: (a) the number of parameters to be estimated affects the number of stimuli that will need to be shown to individuals, (b) attributes should be independent of each other to the extent possible, (c) the conjoint results should be relevant to patients and actionable.

The attributes and levels from Figure 8 form the basis of a conjoint study. A set of full-profile attribute "bundles", four of which are depicted in Figure 9, are created using a design of experiments approach using software. For the purpose of accurately estimating the underlying preferences of each individual patient, the design software maximizes two objectives: balance (each attribute level appearing with roughly equal frequency on the conjoint stimuli) and orthogonality (no correlations between attribute levels consistently appearing together).



73%	Treatment 1	Treatment 2	Treatment 3	Treatment 4
	Doctor and Family Support this treatment Sex: Decreased compared to before treatment	Doctor and Family Support this treatment Sex: Decreased compared to before treatment	Doctor and Family do not favor this treatment Sex: Unable to engage in sex	Doctor and Family Support this treatment Sex: Unable to engage in sex
	Cautious: Treatment gives me months or longer to decide Urinary: Short-term issues	Active: Treatment requires action within weeks Urinary: No problems	Cautious: Treatment gives me months or longer to decide Urinary: No problems	Cautious: Treatment gives me months or longer to decide Urinary: Short-term issues
	No Cutting: Treatment does NOT require any surgery Bowel: No problems	Cutting: Surgery with some risks and hospital time Bowel: No problems	No Cutting: Treatment does NOT require any surgery Bowel: Short term urgent & frequent bowel movements	Cutting: Surgery with some risks and hospital time Bowel: No problems
	Lifespan: Live my expected lifespan	Lifespan: Live 5 years fewer than expected	Lifespan: Live 5 years fewer than expected	Lifespan: Live 5 years fewer than expected
	2	16	11	15
Best	○	☑	○	○
Worst	○	○	☒	○

Figure 9: Conjoint Stimuli and the Best/Worst task in Excel

KEY INITIAL RESULTS BASED ON PATIENT AND DOCTOR RESPONSES:

- **Face Validity**: The seven attributes appear to be reasonably complete in describing patient considerations for prostate cancer treatment. They are perceived to be independent of each other and sufficiently clear.
- **Patient and Doctor Reaction**: Initial surveys of doctors and patients, a sample of which appears below in Figure 10, suggest that the conjoint attributes developed through the VoP methodology are reasonably easy to understand and allow patients to express their treatment-related priorities. Further, both doctors and patients agree that communication between them has been facilitated by discussing the conjoint analysis report generated based on these attributes and levels.
- **Hidden attributes**: The VoP method enabled the research team to confirm obvious treatment-related attributes such as sexual, urinary and bowel function as well as life span. But the process also revealed less obvious conjoint attributes such as the value of

others' opinions, the need to preserve body integrity and avoid surgery, and the desire to take either a quick & active approach to treatment or a more deliberate & cautious one.

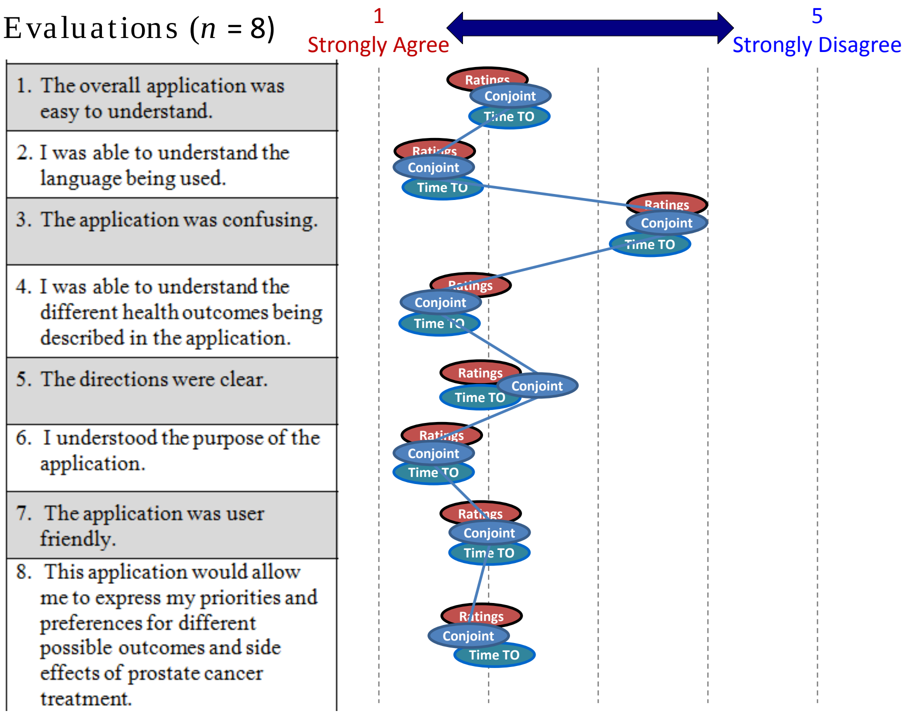


Figure 10: Focus Group Survey Results

LIMITATIONS:

The seven steps from Figure 1 that comprise the voice-of-the-patient process of developing conjoint attributes seems effective based on the final results, but several limitations of this research and opportunities for improvement are evident. The sample size was quite limited, and it is possible that additional attributes might have been discovered with even more interviews. Several steps, most notably the process of identifying the 15 themes and narrowing from almost 1,000 parsed quotations to 70, are rather subjective. One solution would be to employ two or more groups of independent judges to test the level of convergence regarding themes and the final set of quotations. An obvious limitation is the expense and time required to implement the method, which can be considerable. A "gut feel" approach of the researchers would be likely to identify half or more of the attributes discovered through a more rigorous, but more expensive process. Finally, the choice of agglomerative hierarchical clustering (AHC) applied to the simple metric of 0-1 "distance" characterizing affinity between quotations is subject to debate and further consideration. Increasing the "resolution" of affinity distances and/or employing other statistical techniques for grouping quotations should be explored.

DISCUSSION:

The present research and positive results suggests that listening carefully and objectively to patients is key to developing a complete and easily understood set of conjoint attributes and levels. Patient interviews, which have traditionally been viewed as qualitative in nature and have served mainly as guides for the researchers can be parsed, quantified and aggregated in such a way as to let the conjoint attributes “emerge” from the data.

As conjoint analysis in health grows in importance, and diffuses to additional medical applications, the need to carefully develop the attributes and levels used in each study will only increase. The present approach provides a starting point for addressing this need. While it requires a combination of subjective judgment and objective analysis, our empirical application demonstrates that these requirements are not too onerous. We are just at the beginning of the next phase of patient-centered preference measurement, a development that can only improve the all-important quality of communication between health care professionals and their patients.

Bibliography

Burchill, Gary and Hepner-Brodie, Christina (2005), *Voices Into Choices,* Madison, WI: Oriel, Inc.

Griffin, Abbie and Hauser, John R. (1993), “The Voice of the Customer,” *Marketing Science,* 12:1, Winter 1993, pp. 1-27.

An Overview of the Design of Stated Choice Experiments

Warren F. Kuhfeld
SAS Institute Inc.
John C. Wurst
The University of Georgia,
and Atlanta Marketing Sciences Consulting, Inc.

Introduction

Stated choice methods have become a popular approach for researchers to better understand and predict consumer choice behavior since their inception from the seminal work of Louviere and Woodworth (Louviere and Woodworth 1983). The purpose of this paper is to review some of the current methodologies used by practitioners to create stated choice model designs. The discussion begins with a review of choice design fundamentals and issues that arise in design applications. A classification scheme is presented next categorizing the basic approaches to choice design creation. The bulk of the discussion follows presenting a variety of examples illustrating different design creation strategies.

Respondent Utility and Choice

Product and service offerings can be viewed as collections of constituent elements that provide utility to consumers, such as brand name, features, and price. Choice methods attempt to model behavior by estimating the associated consumer utility models that are comprised of the utilities of the component parts. In practice, these utility models are often viewed as either generic, alternative specific, or having a mother logit model structure depending upon the type of effects included.

To illustrate the different types of models, consider a bread maker product example with three attributes (brand, loaf shape, and price), each attribute having two levels as follows.

Brand	Loaf Shape	Price
Oster	Square	\$139
Braun	Rectangular	\$159

Generic Model. A generic model consists of only main effects. Using βs to symbolize attribute level part worth utilities, we have the following for our bread maker example:

$$\beta_{\text{Oster}} \quad \beta_{\text{Braun}} \quad \beta_{\text{sq}} \quad \beta_{\text{rec}} \quad \beta_{\$139} \quad \beta_{\$159}$$

The total utility of any product would be represented by including the appropriate level from each attribute. For example, the utility of a Braun product making square shaped bread priced at \$139 would be: $U = \beta_{\text{Braun}} + \beta_{\text{sq}} + \beta_{\$139}$

Alternative Specific Effects Model. As the name indicates, alternative specific models have utility terms specific to different product alternatives. In the following alternative specific examples, the first represents the total utility of an Oster product that makes square shaped bread priced at $139, and the second specifies the total utility of a Braun product that makes square shaped bread priced at $139.

$$U_{Oster} = \beta_{Oster} + \beta_{Oster(sq)} + \beta_{Oster(\$139)}$$
$$U_{Braun} = \beta_{Braun} + \beta_{Braun(sq)} + \beta_{Braun(\$139)}$$

Note that the part-worth shape and price utilities are specific to the different products (brand feature and brand price interactions).

Mother Logit Model. This model extends the alternative specific model by the addition of cross-effect terms. Cross effects model the impact of a different product's features and price on a particular product. Extending the above example by adding cross-effect terms produces the following:

$$U_{Oster} = \beta_{Oster} + \beta_{Oster(sq)} + \beta_{Oster(\$139)} + \beta_{Oster*Braun(sq)} + \beta_{Oster*Braun(\$139)}$$
$$U_{Braun} = \beta_{Braun} + \beta_{Braun(sq)} + \beta_{Braun(\$139)} + \beta_{Braun*Oster(sq)} + \beta_{Braun*Oster(\$139)}$$

Cross effects are used to address violations of the IIA property of multinomial logit probabilities.

CHOICE DESIGN AND DESIGN EFFICIENCY

The typical respondent task in a choice study involves making selections from sets of alternative product profiles, and often a "none of these" alternative. Profiles are product characterizations obtained by combining levels from the different study attributes. The following is an example set from the bread maker example after adding some additional brand levels to the brand attribute (Sears and Panasonic), and an additional price level ($149).

	Oster	**Braun**	**Sears**	**Panasonic**	**None**
	Square Loaf	**Square Loaf**	**Rectangular Loaf**	**Square Loaf**	
	$149	**$149**	**$159**	**$139**	
Which, if any, would you purchase?	□	□	□	□	□

A choice design is the entire collection of choice sets to be evaluated.

The following is an example of an 18 set design for the bread maker example.

	Oster		Braun		Sears		Panasonic	
Set	Price	Shape	Price	Shape	Price	Shape	Price	Shape
1	$149	Rect.	$149	Square	$149	Square	$139	Rect.
2	$149	Rect.	$139	Square	$159	Rect.	$149	Square
3	$139	Square	$139	Rect.	$139	Rect.	$139	Square
4	$149	Square	$149	Rect.	$149	Rect.	$139	Square
5	$159	Rect.	$159	Rect.	$159	Rect.	$139	Rect.
6	$159	Square	$159	Square	$159	Square	$139	Square
7	$139	Rect.	$159	Rect.	$149	Square	$149	Square
8	$159	Square	$139	Square	$149	Rect.	$159	Rect.
9	$139	Square	$159	Square	$149	Rect.	$149	Rect.
10	$149	Rect.	$159	Rect.	$139	Rect.	$159	Rect.
11	$149	Square	$159	Square	$139	Square	$159	Square
12	$149	Square	$139	Rect.	$159	Square	$149	Rect.
13	$139	Square	$149	Square	$159	Square	$159	Square
14	$159	Rect.	$149	Square	$139	Rect.	$149	Square
15	$139	Rect.	$149	Rect.	$159	Rect.	$159	Rect.
16	$159	Square	$149	Rect.	$139	Square	$149	Rect.
17	$139	Rect.	$139	Square	$139	Square	$139	Rect.
18	$159	Rect.	$139	Rect.	$149	Square	$159	Square

A fundamental goal in choice design development is to create a design that will result in the most precise utility model estimates, in other words, the smallest possible standard errors of the utility estimates. A summary measure of design quality often used in practice is D-efficiency, which is obtained by taking the reciprocal of the *p*th root of the utility estimate variance covariance matrix determinant.

$$D\text{-}efficiency \;=\; \left|V(\widehat{\beta})\right|^{-\left(\frac{1}{p}\right)}$$

Where p = the number or parameters.

Note that the “smaller” the variance matrix, the larger the D-efficiency measure.

While studies commonly model choice probabilities using multinomial logit (McFadden 1974), early choice design work used linear model design principles as a surrogate (Lazari and Anderson 1994; Kuhfeld, Tobias, and Garratt 1994). An attractive feature of the linear model approach is that the D-efficiency measure can be scaled from 0 to 100 with the following relatively simple formulation:

$$D\text{-}efficiency \;=\; 100/N_D\left|(\mathbf{X}'\mathbf{X})^{-1}\right|^{\left(\frac{1}{p}\right)}$$

Where N_D = the number of sets and **X** is the coded design matrix.

More recent choice design work is typically based on the D-efficiency measure utilizing the utility estimate variances from the multinomial logit model:

$$V(\hat{\beta}) = \left[\sum_{k=1}^{n} N \left[\frac{\sum_{j=1}^{m} \exp(x_j'\beta) x_j x_j'}{\sum_{j=1}^{m} \exp(x_j'\beta)} - \frac{\left(\sum_{j=1}^{m} \exp(x_j'\beta) x_j\right)\left(\sum_{j=1}^{m} \exp(x_j'\beta) x_j\right)'}{\left(\sum_{j=1}^{m} \exp(x_j'\beta)\right)^2} \right] \right]^{-1}$$

Where m = number of brands, n = number of sets, N = number of respondents, and x_j = design codes. Note that a prior specification of the utilities (parameters) must be specified to determine the variances, and therefore the associated D-efficiency measure.

Huber and Zwerina (1996) identify four fundamental principles of efficient choice designs: orthogonality, level balance, utility balance, and minimal overlap. Orthogonality pertains to the design attribute levels being independent. Level balance is achieved when for all attributes, all levels within an attribute appear in the design with equal frequency. Utility balance pertains to all alternatives in a choice set having the same total utility. Minimal overlap means there is no repetition of attribute levels across alternatives in a choice set. If all four principles are satisfied, the design is optimal and maximum D-efficiency has been attained.

Choice Design Issues in Practice

While it is possible to create optimal choice designs, they are not commonly found in marketing research applications due to practical considerations. Frequently there is a need to restrict the design in some way. For example, in a financial services credit card study the client won't permit an offering with the lowest fee to appear with the most generous cash back reward. Such restrictions will necessarily diminish design efficiency (except in certain special cases, such as when restrictions are incorporated by creating super attributes consisting of combined levels of the original attributes).

Another common issue is dominance. For example, consider the following choice set for a credit card application:

Attributes	Alternative 1	Alternative 2	Alternative 3
Fee	$29	$59	$99
APR	10%	12%	14%
Cash Back Reward	4%	3%	2%
Which one do you most prefer ?	❑	❑	❑

Note that there is no need to ask this question in the respondent task since it can be determined a priori that rational consumers would prefer the first alternative. While dominance is counter to the utility balance principle, it does occur with design creation methods used in practice and can be addressed by employing restrictions among the alternatives.

Recall the minimal overlap principle of optimal designs. This is consistent with a compensatory decision process that is assumed in standard choice methods (an unattractive feature of one attribute can possibly be compensated by an attractive feature of another). However, research has shown that decision processes are not always compensatory (Gilbride and Allenby 2004; Hauser, Dahan, Yee, and Orlin 2006). Consider the two choice sets below for a credit card study.

Attributes	Alternative 1	Alternative 2	Alternative 3
Brand	Am.Ex.	Cap. One	Chase
Fee	$99	$59	$29
APR	10%	12%	14%
Cash Back Reward	2%	4%	3%
Which one do you most prefer ?	□	□	□

Attributes	Alternative 1	Alternative 2	Alternative 3
Brand	Am.Ex.	Cap. One	Cap. One
Fee	$99	$59	$29
APR	10%	12%	14%
Cash Back Reward	2%	4%	3%
Which one do you most prefer ?	□	□	□

The first has no level overlap, the second has overlap in the brand attribute (The Capital One brand appears in two alternatives). Suppose a respondent is using a non-compensatory decision process where they “must have” a Capital one credit card. With such a decision process, a design without overlap would provide no information about the other attributes. In choice sets with overlap similar to the second one above we would learn how the other attributes affect the respondent’s choices.

Another common design issue facing practitioners is how to successfully administer the respondent task for a study with many attributes without over taxing respondents. Choice tasks require respondents to evaluate multiple alternatives in each task, and after about 7 or so attributes (depending on the attribute complexity) the task can become burdensome. To address this, Chrzan and Elrod (1995) developed the partial profile approach to choice designs. Partial profile designs greatly reduce respondent burden by only presenting a subset of the study attributes in any one choice set. The following example presents full and partial profile choice sets from a printer study with 11 attributes.

Full Profile

Attributes	Alternative 1	Alternative 2	Alternative 3
Brand	HP	Kodak	Canon
Type	Basic	Premium	Docking
Wireless Printing	Yes	No	Yes
Memory Card Slots	Yes	Yes	No
Zoom and Crop	Manual	Auto	Auto
Print Speed	15 sec.	45 sec.	30 sec.
Computer free printing	No	Yes	Yes
Print Dry Time	Instant	15 sec.	30 sec.
Auto red eye removal	No	Yes	Yes
Photo Life	25 years	50 years	100 years
Price	$79	$99	$119

Partial Profile

Attributes	Alternative 1	Alternative 2	Alternative 3
Brand	HP	Kodak	Canon
Type	Basic	Premium	Docking
Zoom and Crop	Manual	Auto	Auto

MAXIMUM DIFFERENCE SCALING (MAXDIFF): A SPECIAL CASE

MaxDiff is a popular methodology, developed by Louviere (Louviere 1991, Finn and Louviere 1992) that produces item scale values conveying both order and magnitude based on simple comparative judgments. Respondents are presented with sets of items, typically 3 to 6 in a set, and choose the most preferred and the least preferred in the set (or the most important, least important; best item, worst item, etc.).   An example MaxDiff set appears below for a toothpaste study involving seven total tooth paste items/attributes.

How Important are attributes when you purchase toothpaste?

Of these four, which are most and least important?

Most Important		Least Important
❑	**tarter control**	❑
❑	**whitening**	❑
❑	**flavor**	❑
❑	**breath freshening**	❑

Note that MaxDiff is a type of choice study where two choices are obtained for each set. Choice design principles and analysis methods are used to create the respondent task and analyze results. MaxDiff designs fall in the category of balanced and partially balanced incomplete block designs.

A CLASSIFICATION OF DESIGN CREATION APPROACHES

Design creation approaches can be categorized into three basic classes: combinatorial methods, structurally directed approaches, and efficiency directed approaches. The earliest procedures were combinatorial.   These methods develop designs mathematically resulting in only optimal designs (orthogonality, level balance, utility balance, and minimal overlap fully achieved). Designs created using this approach can be found in design catalogs.   However, this approach alone is not common in marketing research practice partly due to the relatively limited design configurations that are permissible.

Structurally directed methods develop designs through computer iterations with the objective of satisfying certain specified design characteristics, such as at least some of the four design principles (orthogonality, level balance, utility balance, and minimal overlap). Note that fully satisfying any of the specified characteristics is not a requirement, only that such procedures work toward achieving the objective. Examples include the randomized and adaptive strategies of Sawtooth Software.

Efficiency directed methods create designs through computer iterations with the objective of maximizing an efficiency measure. As with the structurally directed approaches, fully achieving the objective is not required, only that the methods work toward achieving the objective. Examples include algorithms that maximize D-efficiency, such as those found in SAS.

Unlike combinatorial methods, both the structurally directed and efficiency directed approaches are able to accommodate restrictions and other practical considerations faced by researchers that diminish design efficiency. Practical considerations may run counter to statistical design efficiency, but can result in better quality designs than methods based only on statistical criteria.

Examples

The rest of this paper goes through a number of examples. For a series of choice models, we construct 100% efficient, optimal designs by directly using combinatorial construction methods. Other examples show how computerized searches fare in the same problem. In most real-life applications, direct combinatorial construction methods are too limiting, so you must use other methods such as computerized searches. By comparing both optimal and nonoptimal results, you can develop an understanding of the properties of efficient choice designs.

The assumption that $\boldsymbol{\beta} = \mathbf{0}$ is repeatedly emphasized throughout these examples. You can make other assumptions, and your design efficiency will depend on this assumption. Designs constructed by combinatorial means are only guaranteed to be optimal when $\boldsymbol{\beta} = \mathbf{0}$.

Our examples show SAS code for completeness and to be fully explicit about the construction methods, but we do not discuss SAS syntax. These examples focus on optimal choice designs, nonoptimal but efficient choice designs, and understanding their properties. The focus is not on software. More details on the software can be found in Kuhfeld (2010a).

Combinatorial Approach to Main-Effects Choice Designs. Consider a simple choice model with 4 two-level attributes. The goal is to construct a choice design with four choice sets and two alternatives per set. A main-effects model is fit, so there are no interactions, alternative-specific effects, or other higher order effects. The goal is to construct an optimal design for this model under the assumption that the parameter vector is all zero. This design can easily be constructed from the orthogonal array $4^1 2^4$. Orthogonal arrays are main-effects plans that were first developed for linear models. Orthogonal array construction methods are beyond the scope of this paper. However, you can see Kuhfeld (2010ab) for more information. In short, many orthogonal arrays are constructed using combinatorial recipes which involve specialized methods using integer field arithmetic.

The procedure for constructing the choice design from an orthogonal array is simple. The four-level factor is used as the choice set number, and the remaining factors provide the attributes for the choice design. In the SAS system, you can use the MktEx macro to create over 117,000 different orthogonal arrays. Example:

```
%mktex(4 2 ** 4, n=8)
```

Two different versions of the design are created, one that is sorted and uses the original level assignment, and one that is randomized (the rows are sorted into a random order, and levels are randomly reassigned). The randomized design is more suitable for use, so we will use it and sort the rows by choice set number.

```
proc sort data=randomized; by x1; run;
```

Next, we will reassign the variable names and display the results.

```
%mktlab(data=randomized, out=chdes, vars=Set x1-x4)
proc print; id set; by set; run;
```

The design is as follows:

```
Set    x1    x2    x3    x4

 1      2     1     2     2
        1     2     1     1

 2      1     1     1     2
        2     2     2     1

 3      2     1     1     1
        1     2     2     2

 4      1     1     2     1
        2     2     1     2
```

The first thing to do with any candidate choice design is evaluate its efficiency and the standard errors and variances. Particularly for small designs, you should examine the covariances as well.

```
%choiceff(data=chdes, init=chdes(keep=set),  options=relative, nalts=2, nsets=4, model=class(x1-x4 / standorth), beta=zero)
proc print data=bestcov noobs; id __label; var x:; label __label='00'x; run;
```

The ChoicEff macro uses a main-effects model, an assumed parameter vector of zero, and a standardized orthogonal contrast coding to evaluate the design. With this coding and this type of model, the maximum possible *D*-efficiency is the number of choice sets. The first table gives the design efficiency.

```
          Final Results

Design                      1
Choice Sets                 4
Alternatives                2
Parameters                  4
Maximum Parameters          4
D-Efficiency           4.0000
Relative D-Eff       100.0000
D-Error                0.2500
1 / Choice Sets        0.2500
```

Relative to a choice design with a raw *D*-efficiency of 4, which is the number of choice sets, this design is 100% efficient. There are four parameters, and the maximum number of parameters with four choice sets and two alternatives is four. Usually, you would hope to have extra choice sets that provide enough information for at least the potential of additional parameters. However, designs like this one are saturated (also known as tight) and provide the maximum number of parameters. The standard errors and variances are as follows:

```
      Variable                                       Standard
n        Name       Label     Variance     DF           Error

1        x11        x1 1         0.25       1             0.5
2        x21        x2 1         0.25       1             0.5
3        x31        x3 1         0.25       1             0.5
4        x41        x4 1         0.25       1             0.5
                                           ==
                                            4
```

The variances and standard errors are constant, and all four main-effects parameters can be estimated with equal precision. This and every example assumes a sample size of one. The actual variances are multiplied by 1/*N* with *N* subjects. You can specify *N* in the ChoicEff macro with the **n=** option. The point in all examples is to compare relative variances not the actual values. The variances and covariances are as follows:

```
              x1 1     x2 1      x3 1     x4 1

     x1 1     0.25     0.00     -0.00     0.00
     x2 1     0.00     0.25      0.00     0.00
     x3 1    -0.00     0.00      0.25     0.00
     x4 1     0.00     0.00      0.00     0.25
```

All covariances are within rounding error of zero. This design is 100% *D*-efficient and hence optimal for a main-effects only choice model under the assumption that $\boldsymbol{\beta} = \mathbf{0}$. It is not guaranteed to be optimal for any other circumstances such as for models with interactions or parameter vectors that are not zero. In general, orthogonal arrays, where one factor provides the choice set number, provide a way to get optimal designs for main-effects choice models under the assumption that $\boldsymbol{\beta} = \mathbf{0}$. They do not exist for many reasonable choice models, but when they exist, they are optimal. However, it must be emphasized that they are optimal for main-effects only choice models with $\boldsymbol{\beta} = \mathbf{0}$. Orthogonal arrays, that can be used to make optimal choice designs (given as design followed by the total number of alternatives) include: 2^4 4^1, 8; 3^4, 9; 2^8 8^1, 16; 4^5, 16; 3^6 6^1, 18; 2^{12} 12^1, 24; 5^6, 25; 3^9 9^1, 27; 2^{16} 16^1, 32; 4^8 8^1, 32; 3^{12} 12^1, 36; 2^{20} 20^1, 40; 3^9 15^1, 45; 2^{24} 24^1, 48; 4^{12} 12^1, 48; 7^8, 49; 5^{10} 10^1, 50; 3^{18} 18^1, 54; 2^{28} 28^1, 56; 3^{12} 21^1, 63; 2^{32} 32^1, 64; 4^{16} 16^1, 64; 8^9, 64; 2^{36} 36^1, 72; 3^{24} 24^1, 72; 5^8 15^1, 75; 2^{40} 40^1, 80; 4^{10} 20^1, 80; 3^{27} 27^1, 81; 9^{10}, 81; 2^{44} 44^1, 88; 3^{30} 30^1, 90; 2^{48} 48^1, 96; 7^{14} 14^1, 98; 3^{13} 33^1, 99; 5^{20} 20^1, 100; 10^4, 100; and so on. The factor with the most levels is used as the choice set number. An orthogonal array with *n* total alternatives and an *m*-level factor produces a choice design *m* choice sets and *n* / *m* alternatives per choice set. You do not need SAS software to make these designs. You can instead go to the SAS web site (Kuhfeld 2010b), which freely provides orthogonal arrays.

For this design and for designs like the ones previously listed, the optimal design has a variance matrix that is diagonal. In fact it is a constant (one over the number of choice sets) times the identity matrix. This clearly shows that this design is optimal. For most real-world problems, you will not get diagonal variance matrices. In fact, for most real-world problems, even the optimal design (if you knew it) would not have a diagonal variance matrix. These optimal combinatorial designs are very symmetric. All attributes have the same number of levels, and that number is the number of alternatives in each choice set. In practice, most designs are asymmetric and have varying numbers of levels in the attributes that do not match the number of alternatives. Asymmetry virtually always precludes having a diagonal variance matrix.

Some researchers (e.g. Street and Burgess 2007) advocate designs with these properties (although they approach the construction process differently). Others (e.g. Chrzan, Zepp, and White 2010), caution that these designs have negative traits in spite of their statistical optimality. Each researcher needs to make his or her own decision based on the needs of the study. Certainly these designs are useful from the point of view of understanding basic choice design concepts and providing optimal benchmarks to which other designs can be compared.

Efficiency Directed Search for Main-Effects Choice Designs. In the preceding section, we saw an example of a situation where an optimal design can be constructed directly from an orthogonal array. We also saw lists of some of the small orthogonal arrays that lend themselves to choice design construction. These designs tend to be very symmetric. In practice, the researcher might find that level of symmetry unrealistic, or might not be willing to assume that $\boldsymbol{\beta} = \boldsymbol{0}$. In those cases, different approaches are used. In one commonly used approach, a computerized search is performed where an efficient design is constructed from a set of candidate alternatives.

The researcher starts by determining a choice model, the number of alternatives, number of levels for each alternative, number of choice sets, and an assumed parameter vector. The researcher then creates the candidate set of alternatives with all of the right attributes and levels. The algorithm starts by creating a design from a random set of candidates. It iteratively evaluates the effect of removing an alternative and replacing it by each candidate. Swaps that improve efficiency are retained, and the rest are discarded. Iteration ceases when efficiency quits improving. This process is repeated with different random initial designs. The most efficient design is chosen. For most reasonable design specifications (that is, for designs that are not too small or are overly restricted), you would expect the design to be highly efficient. Equivalently, you expect all parameters to be estimable with reasonably small standard errors.

Consider again the problem from the preceding example: a choice design with four choice sets, two alternatives per set, 4 two-level attributes, a main-effects model, and $\boldsymbol{\beta} = \boldsymbol{0}$. This is a very small problem as choice designs go, yet there are over 100 million possible choice designs. A computerized search will never consider them all, but it can still do a good job of finding the optimal design for a problem this size. In fact, it finds the optimal design for over 80% of the random initializations. This success rate rapidly decreases as design size increases. The following steps create a full-factorial candidate set of alternatives, search for an efficient choice design for a main-effects choice model with $\boldsymbol{\beta} = \boldsymbol{0}$, and display the results:

```
%mktex(2 ** 4, n=16, seed=17)

%choiceff(data=randomized, model=class(x1-x4 / standorth), beta=zero,  seed=151, options=relative, flags=2, nsets=4)

proc print data=best; id set; by set; var x:; run;
```

The result is a 100% D-efficient choice design.

```
Set    x1    x2    x3    x4

 1      2     2     2     2
        1     1     1     1

 2      1     1     2     2
        2     2     1     1

 3      2     1     2     1
        1     2     1     2

 4      1     2     2     1
        2     1     1     2
```

Now consider a different problem, one that is asymmetric. A researcher is interested in constructing a choice design for a main-effects only model, with $\boldsymbol{\beta} = \boldsymbol{0}$, 4 two-level attributes and 4 three-level attributes, 18 choice sets, and three alternatives.

```
%mktex(2 ** 4 3 ** 4, n=1296, seed=17)

%choiceff(data=randomized, model=class(x1-x8 / standorth), beta=zero,
     seed=151, options=relative, flags=3, nsets=18, maxiter=10)

proc print data=best; id set; by set; var x:; run;
```

We know that with an asymmetric design like this that the covariance matrix will not be diagonal. We also know that all of our variances will not equal one over the number of choice sets. They will be bigger.

Some of the results are as follows:

```
                          Final Results

                    Design                     6
                    Choice Sets               18
                    Alternatives               3
                    Parameters                12
                    Maximum Parameters        36
                    D-Efficiency         17.1319
                    Relative D-Eff       95.1775
                    D-Error               0.0584
                    1 / Choice Sets       0.0556

          Variable                                         Standard
   n        Name       Label     Variance      DF          Error

   1        x11        x1 1      0.063182       1        0.25136
   2        x21        x2 1      0.063143       1        0.25128
   3        x31        x3 1      0.063183       1        0.25136
   4        x41        x4 1      0.063252       1        0.25150
   5        x51        x5 1      0.057129       1        0.23902
   6        x52        x5 2      0.056085       1        0.23682
   7        x61        x6 1      0.058072       1        0.24098
   8        x62        x6 2      0.056093       1        0.23684
   9        x71        x7 1      0.058012       1        0.24086
  10        x72        x7 2      0.057348       1        0.23948
  11        x81        x8 1      0.056951       1        0.23864
  12        x82        x8 2      0.056093       1        0.23684
                                               ==
                                               12
```

These results look quite good. We see that we have 12 parameters and that we can estimate at most 36 parameters. In most cases, you want to have enough choice sets to estimate more parameters than you will actually estimate. One over the number of choice sets is 0.0556. None of the variances are that small, but all are close. The variances for x1-x4 (the two-level attributes) are greater than the variances for x5-x8 (the three-level factors). The number of levels does not evenly divide the number of alternatives with the two-level factors, so balance is not possible within choice set, and the variances are inflated. The *D*-efficiency, relative to a design with a variance matrix of an identity matrix I times one over the number of choice sets is 95.1775%. Relative to this hypothetical optimal design, our design is 95.1775% *D*-efficient. Relative to the true and unknown optimal design, our design has some unknown *D*-efficiency, but it must be greater than 95.1775%.

Such is the art of choice design. We know that the search algorithms find optimal designs for small problems. We know by comparing variances and computing *D*-efficiency relative to hypothetical optimal *D*-efficiency values that our designs are good. However, in practice, we typically do not know precisely how good our design is. For larger and more complicated problems and designs with restrictions, we often do not know a reasonable maximum efficiency. We always know the variances, and examining them for consistency and overly large values is important.

Combinatorial Approach to Alternative-Specific Effects Choice Designs. We will consider in this example a problem where there are three alternatives that have labels such as brand names. Each choice set must contain all three brands. We will create a design for 27 choice sets, three alternatives per set, four three-level attributes, an alternative-specific effects model, and ***ß = 0.*** We will construct this design from an orthogonal array that has 12 factors (three alternatives times four attributes) and 27 rows (one per choice set).

The following steps create the orthogonal array, convert it from one row per choice set to one row per alternative, add the brand names, display the first three choice sets, and evaluate the design:

```
%mktex(3 ** 12, n=27, seed=104)

%mktkey(3 4)

data key; Brand = scan('A B C', _n_); set key; run;

%mktroll(design=randomized, key=key, out=chdes, alt=brand)

proc print data=chdes(obs=9); by set; id set brand; var x:; run;

%choiceff(data=chdes, init=chdes(keep=set),  nalts=3, nsets=27, rscale=alt, beta=zero,
     model=class(brand / sta)  class(brand * x1 brand * x2 brand * x3 brand * x4 / sta zero=' '))

proc print data=bestcov noobs; id __label; var B:; label __label='00'x; run;
```

The first three choice sets are as follows:

```
Set    Brand    x1    x2    x3    x4

 1       A       3     1     3     2
         B       2     1     3     1
         C       3     3     3     2

 2       A       1     1     2     3
         B       1     1     3     2
         C       2     1     3     3

 3       A       3     3     1     2
         B       3     3     3     2
         C       2     1     2     1
```

The design efficiency results are as follows:

```
           Final Results

Design                        1
Choice Sets                  27
Alternatives                  3
Parameters                   26
Maximum Parameters           54
D-Efficiency             6.7359
Relative D-Eff         100.0000
D-Error                  0.1485
1 / Choice Sets          0.0370
```

D-efficiency, relative to an optimal design with an alternative-specific effects model and $\boldsymbol{\beta} = \boldsymbol{0}$, is 100%. The variances and standard errors are as follows:

```
        Variable                                                Standard
   n    Name         Label                Variance     DF          Error

   1    BrandA       Brand A               0.03704      1        0.19245
   2    BrandB       Brand B               0.03704      1        0.19245
   3    BrandAx11    Brand A * x1 1        0.16667      1        0.40825
   4    BrandAx12    Brand A * x1 2        0.16667      1        0.40825
   5    BrandBx11    Brand B * x1 1        0.16667      1        0.40825
   6    BrandBx12    Brand B * x1 2        0.16667      1        0.40825
   7    BrandCx11    Brand C * x1 1        0.16667      1        0.40825
   8    BrandCx12    Brand C * x1 2        0.16667      1        0.40825
   9    BrandAx21    Brand A * x2 1        0.16667      1        0.40825
  10    BrandAx22    Brand A * x2 2        0.16667      1        0.40825
  11    BrandBx21    Brand B * x2 1        0.16667      1        0.40825
  12    BrandBx22    Brand B * x2 2        0.16667      1        0.40825
  13    BrandCx21    Brand C * x2 1        0.16667      1        0.40825
  14    BrandCx22    Brand C * x2 2        0.16667      1        0.40825
  15    BrandAx31    Brand A * x3 1        0.16667      1        0.40825
  16    BrandAx32    Brand A * x3 2        0.16667      1        0.40825
  17    BrandBx31    Brand B * x3 1        0.16667      1        0.40825
  18    BrandBx32    Brand B * x3 2        0.16667      1        0.40825
  19    BrandCx31    Brand C * x3 1        0.16667      1        0.40825
  20    BrandCx32    Brand C * x3 2        0.16667      1        0.40825
  21    BrandAx41    Brand A * x4 1        0.16667      1        0.40825
  22    BrandAx42    Brand A * x4 2        0.16667      1        0.40825
  23    BrandBx41    Brand B * x4 1        0.16667      1        0.40825
  24    BrandBx42    Brand B * x4 2        0.16667      1        0.40825
  25    BrandCx41    Brand C * x4 1        0.16667      1        0.40825
  26    BrandCx42    Brand C * x4 2        0.16667      1        0.40825
                                                       ==
                                                       26
```

All parameters are estimable. The variances for the brand main effects are constant and are one over the number of choice sets. The variances for the alternative-specific effects are constant and are larger. There are many more alternatives corresponding to a particular brand than there are alternatives that correspond to a brand and attribute level combination, so variances will necessarily be smaller for the former and larger for the latter. The variance matrix (not shown) is diagonal with the values in the preceding table on the diagonal. The relative D-efficiency calculations are based on the optimal determinant of a variance matrix with this partitioned structure of main effects and alternative-specific effects. In designs that are not fully symmetric, 100% relative *D*-efficiency will not be achieved, even for optimal designs.

Efficiency Directed Search for Alternative-Specific Effects Choice Designs. Using a computer to search for a choice design with alternative-specific effects is similar to searching for main-effects designs. However, a few details are different. We will consider in this example the same problem as the previous example. There are three alternatives that have labels such as brand names. Each choice set must contain all three brands. We will create a design for 27 choice sets, three alternatives per set, four three-level attributes, an alternative-specific effects model, and $\boldsymbol{\beta} = \boldsymbol{0}$. For this problem, we will need to create a candidate set of alternatives with three parts, one for each brand. The following steps make the candidate points and display a few of them:

```
%mktex(3 ** 5, n=243)

data cand(drop=x5);
  set design;
  Brand = scan("A B C", x5);
  array f[3];
  retain f1-f3 0;
  f[x5] = 1; output; f[x5] = 0;
  run;

proc print noobs data=cand(obs=12); run;
```

The first 12 of the 243 candidates are:

```
x1    x2    x3    x4    Brand    f1    f2    f3

 1     1     1     1      A       1     0     0
 1     1     1     1      B       0     1     0
 1     1     1     1      C       0     0     1
 1     1     1     2      A       1     0     0
 1     1     1     2      B       0     1     0
 1     1     1     2      C       0     0     1
 1     1     1     3      A       1     0     0
 1     1     1     3      B       0     1     0
 1     1     1     3      C       0     0     1
 1     1     2     1      A       1     0     0
 1     1     2     1      B       0     1     0
 1     1     2     1      C       0     0     1
```

In this example, candidates are constructed where all of the three-level factors are the same for each brand. This is not a requirement. In cases of extreme asymmetry, both the attributes and levels can be different for every alternative. No or little asymmetry across alternatives is more typical. The candidate set has “flag variables,” f1-f3, that flag which candidates can be used for alternative *i* (when f[*i*] is 1).

The following step searches the candidates for an efficient design:

```
%choiceff(data=cand,  beta=zero, seed=93, rscale=alt, flags=f1-f3, nsets=27, maxiter=100,
    model=class(brand / sta) class(brand * x1 brand * x2 brand * x3 / standorth zero=''))

proc print data=bestcov noobs; id __label; var B:; label __label='00'x; run;
```

Some of the results are as follows:

```
          Final Results

Design                    16
Choice Sets               27
Alternatives               3
Parameters                20
Maximum Parameters        54
D-Efficiency          6.8612
Relative D-Eff       98.3850
D-Error               0.1457
1 / Choice Sets       0.0370
```

```
         Variable                                                Standard
  n      Name          Label               Variance     DF         Error

  1      BrandA        Brand A              0.03704      1       0.19245
  2      BrandB        Brand B              0.03704      1       0.19245
  3      BrandAx11     Brand A * x1 1       0.17055      1       0.41298
  4      BrandAx12     Brand A * x1 2       0.17196      1       0.41468
  5      BrandBx11     Brand B * x1 1       0.17225      1       0.41503
  6      BrandBx12     Brand B * x1 2       0.17343      1       0.41645
  7      BrandCx11     Brand C * x1 1       0.16944      1       0.41163
  8      BrandCx12     Brand C * x1 2       0.17160      1       0.41424
  9      BrandAx21     Brand A * x2 1       0.17156      1       0.41420
 10      BrandAx22     Brand A * x2 2       0.16828      1       0.41021
 11      BrandBx21     Brand B * x2 1       0.17748      1       0.42128
 12      BrandBx22     Brand B * x2 2       0.17177      1       0.41445
 13      BrandCx21     Brand C * x2 1       0.17525      1       0.41863
 14      BrandCx22     Brand C * x2 2       0.17176      1       0.41444
 15      BrandAx31     Brand A * x3 1       0.17296      1       0.41588
 16      BrandAx32     Brand A * x3 2       0.17781      1       0.42168
 17      BrandBx31     Brand B * x3 1       0.17551      1       0.41893
 18      BrandBx32     Brand B * x3 2       0.17498      1       0.41831
 19      BrandCx31     Brand C * x3 1       0.17258      1       0.41543
 20      BrandCx32     Brand C * x3 2       0.17583      1       0.41932
                                                        ==
                                                        20
```

D-efficiency, relative to an optimal design with an alternative-specific effects model and ***β* = *0*,** is 98.385%. All parameters are estimable. The variances for the brand main effects are constant and are one over the number of choice sets. The variances for the alternative-specific effects are not constant. They are similar to the variances in the preceding example, but these are larger. The variance matrix (not shown) is diagonal for only the first two rows and columns. The relative *D*-efficiency calculations are based on the optimal design, like the one created in the preceding section. This design is close, but it is not there.

The candidate set for this problem is a full-factorial design, and it is reasonably small at 243 rows. However, the model with 20 parameters is much more complex. The computerized search finds with efficiency greater than 98% quickly, easily, and often, but it will rarely find the optimal design for a problem this size or larger.

Random Design for Alternative-Specific Effects Choice Model. Sawtooth Software creates designs that are large and have a substantial random component to them. In this example, we will create a random design for a model with 198 choice sets, three alternatives per set, four three-level attributes, an alternative-specific effects model, and ***β = 0.*** Next we compare that design to a design created by using the random design as a candidate set and to an optimal design. The following steps generate and evaluate the three designs:

```
%let sets = 198;
data random;
  do Set = 1 to &sets;        do Brand = 'A', 'B', 'C';
      x1 = ceil(3 * uniform(292));      x2 = ceil(3 * uniform(292));      x3 = ceil(3 * uniform(292));      x4 = ceil(3 *
uniform(292));
      f1 = brand = 'A';     f2 = brand = 'B';     f3 = brand = 'C';
      output;
      end;        end;
  run;

%mktex(3 ** 12, n=&sets, seed=104)
%mktkey(3 4)              data key; Brand = scan('A B C', _n_); set key; run;
%mktroll(design=randomized, key=key, out=chdes, alt=brand)

%choiceff(data=random, init=random(keep=set), nalts=3, nsets=&sets, rscale=alt, beta=zero,
      model=class(brand / sta)  class(brand * x1 brand * x2 brand * x3 brand * x4 / sta zero=' '))
%choiceff(data=random, flags=f1-f3, nsets=&sets, rscale=alt, beta=zero,
      model=class(brand / sta)  class(brand * x1 brand * x2 brand * x3 brand * x4 / sta zero=' '))
%choiceff(data=chdes, init=chdes(keep=set),  nalts=3, nsets=&sets, rscale=alt, beta=zero,
      model=class(brand / sta)  class(brand * x1 brand * x2 brand * x3 brand * x4 / sta zero=' '))
```

The three variance columns are displayed in a single table along with relative *D*-efficiency:

```
                          98.69%      99.94%       100%
                          Random      Random       Optimal
                          Design      Candidates   Design
 n    Label               Variance    Variance     Variance     DF

 1    Brand A             0.005087    0.005051     0.005051      1
 2    Brand B             0.005091    0.005051     0.005051      1
 3    Brand A * x1 1      0.023501    0.022777     0.022727      1
 4    Brand A * x1 2      0.023258    0.022745     0.022727      1
 5    Brand B * x1 1      0.023116    0.022751     0.022727      1
 6    Brand B * x1 2      0.023385    0.022750     0.022727      1
 7    Brand C * x1 1      0.023640    0.022773     0.022727      1
 8    Brand C * x1 2      0.022951    0.022759     0.022727      1
 9    Brand A * x2 1      0.022713    0.022772     0.022727      1
10    Brand A * x2 2      0.023749    0.022772     0.022727      1
11    Brand B * x2 1      0.023769    0.022764     0.022727      1
12    Brand B * x2 2      0.022852    0.022794     0.022727      1
13    Brand C * x2 1      0.023674    0.022760     0.022727      1
14    Brand C * x2 2      0.023198    0.022780     0.022727      1
15    Brand A * x3 1      0.023690    0.022591     0.022727      1
16    Brand A * x3 2      0.023121    0.022941     0.022727      1
17    Brand B * x3 1      0.022747    0.022744     0.022727      1
18    Brand B * x3 2      0.023796    0.022749     0.022727      1
19    Brand C * x3 1      0.023505    0.022758     0.022727      1
20    Brand C * x3 2      0.023384    0.022754     0.022727      1
21    Brand A * x4 1      0.022795    0.022755     0.022727      1
22    Brand A * x4 2      0.024001    0.022756     0.022727      1
23    Brand B * x4 1      0.024123    0.022751     0.022727      1
24    Brand B * x4 2      0.022811    0.022767     0.022727      1
25    Brand C * x4 1      0.022748    0.022751     0.022727      1
26    Brand C * x4 2      0.024512    0.022773     0.022727      1
                                                                ==
                                                                26
```

All three designs have great D-efficiency and small variances. The variances for the random design are slightly larger and more variable than the variances for the efficiency-directed search of the random candidate set, which are slightly larger and more variable than the constant (within effect type) variances for the optimal design. With randomized designs, as the design size gets large, D-efficiency approaches 100% and variances become very stable. Efficiency directed searches can do better, but there is very little room to improve on large random designs.

Combinatorial Approach to Partial-Profile Choice Designs. Consider a simple choice model with 16 three-level attributes. A main-effects model is fit, so there are no interactions, alternative-specific effects, or other higher order effects. The goal is to construct an optimal design for this model under the assumption that the parameter vector is all zero. There are three alternatives per set, four of 16 attributes varying per set, and 12 of 16 attributes constant in each set. Choice designs where different subsets of attributes vary within a set and all other attributes are constant (effectively omitting them from the set) are called partial profile designs (Chrzan and Elrod 1995). Optimal partial profile designs can be constructed from an orthogonal array and a balanced incomplete block design (BIBD) (Anderson, 2003). The following steps, create, evaluate, and display an optimal partial profile design with 120 choice sets.

```
%mktbibd(b=20, nattrs=16, setsize=4, seed=17)
%mktex(6 3**4, n=18, seed=151)
proc sort data=randomized out=randes(drop=x1); by x2 x1; run;
%mktppro(design=randes, ibd=bibd)
%choiceff(data=chdes, init=chdes(keep=set), model=class(x1-x16 / sta), nsets=120, nalts=3, rscale=partial=4 of 16, beta=zero)
proc print data=bestcov noobs; id __label; var x:; label __label='00'x; run;
proc print; id set; by set; var x:; run;
```

The following is a BIBD:

```
 x1    x2    x3    x4

  1     4    16    10
 16    13    14    11
 12    15     4    14
  8    10     2    13
 13     9     1    15
  4     6    13     5
 10     5    15     7
  5    12     8    16
  9    14     5     2
  3     5    11     1
  7    13     3    12
 16     2     3    15
 14     6    10     3
  8     3     9     4
  9    11    12    10
  1     7    14     8
  2     1    12     6
 11     4     2     7
  6    16     7     9
 15     8     6    11
```

Interpreting the first row, in the first set of choice sets, attributes 1, 4, 16, and 10 will vary while the other attributes are constant. Interpreting the last row, in the last set of choice sets, attributes 15, 8, 6, and 11 will vary while the other attributes are constant.

The following are some deliberately chosen rows of an orthogonal array:

```
x2    x3    x4    x5

 1     1     2     3
 1     2     3     1
 1     1     3     2
 1     2     1     3
 1     3     2     2
 1     3     1     1

 2     3     3     1
 2     1     1     2
 2     3     1     3
 2     1     2     1
 2     2     3     3
 2     2     2     2

 3     2     1     2
 3     3     2     3
 3     2     2     1
 3     3     3     2
 3     1     1     1
 3     1     3     3
```

The array $3^6 6^1$ is used, sorting by a three-level factor and a six-level factor (x2), but discarding the six level factor. The goal is to ensure that x2=1 rows are rows 1, 7, and 13, x2=2 rows are rows 2, 8, and 14, and so on. The orthogonal array provides the attribute levels, and the BIBD provides the numbers of the attributes that vary. A few of the resulting choice sets are as follows:

```
Set  x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 x15 x16

  1   1  1  1  1  1  1  1  1  1  3   1   1   1   1   1   2
      2  1  1  3  1  1  1  1  1  1   1   1   1   1   1   3
      3  1  1  2  1  1  1  1  1  2   1   1   1   1   1   1

  2   1  1  1  2  1  1  1  1  1  1   1   1   1   1   1   3
      2  1  1  1  1  1  1  1  1  2   1   1   1   1   1   1
      3  1  1  3  1  1  1  1  1  3   1   1   1   1   1   2

 61   1  1  2  1  1  1  1  1  1  1   1   3   1   1   1   1
      1  1  3  1  1  1  2  1  1  1   1   1   3   1   1   1
      1  1  1  1  1  1  3  1  1  1   1   2   2   1   1   1

 91   1  1  1  1  1  1  1  3  1  1   1   1   1   2   1   1
      2  1  1  1  1  1  3  1  1  1   1   1   1   3   1   1
      3  1  1  1  1  1  2  2  1  1   1   1   1   1   1   1

120   1  1  1  1  1  1  1  3  1  1   1   1   1   1   1   1
      1  1  1  1  1  2  1  2  1  1   2   1   1   1   2   1
      1  1  1  1  1  3  1  1  1  1   3   1   1   1   3   1
```

The choice design evaluation results are as follows:

```
          Final Results

Design                       1
Choice Sets                120
Alternatives                 3
Parameters                  32
Maximum Parameters         240
D-Efficiency           30.0000
Relative D-Eff        100.0000
D-Error                 0.0333
1 / Choice Sets       0.008333
```

```
         Variable                                            Standard
   n       Name        Label      Variance      DF           Error

   1       x11         x1 1       0.033333       1          0.18257
   2       x12         x1 2       0.033333       1          0.18257
   3       x21         x2 1       0.033333       1          0.18257
   4       x22         x2 2       0.033333       1          0.18257
   5       x31         x3 1       0.033333       1          0.18257
   6       x32         x3 2       0.033333       1          0.18257
   7       x41         x4 1       0.033333       1          0.18257
   8       x42         x4 2       0.033333       1          0.18257
   9       x51         x5 1       0.033333       1          0.18257
  10       x52         x5 2       0.033333       1          0.18257
  11       x61         x6 1       0.033333       1          0.18257
  12       x62         x6 2       0.033333       1          0.18257
  13       x71         x7 1       0.033333       1          0.18257
  14       x72         x7 2       0.033333       1          0.18257
  15       x81         x8 1       0.033333       1          0.18257
  16       x82         x8 2       0.033333       1          0.18257
  17       x91         x9 1       0.033333       1          0.18257
  18       x92         x9 2       0.033333       1          0.18257
  19       x101        x10 1      0.033333       1          0.18257
  20       x102        x10 2      0.033333       1          0.18257
  21       x111        x11 1      0.033333       1          0.18257
  22       x112        x11 2      0.033333       1          0.18257
  23       x121        x12 1      0.033333       1          0.18257
  24       x122        x12 2      0.033333       1          0.18257
  25       x131        x13 1      0.033333       1          0.18257
  26       x132        x13 2      0.033333       1          0.18257
  27       x141        x14 1      0.033333       1          0.18257
  28       x142        x14 2      0.033333       1          0.18257
  29       x151        x15 1      0.033333       1          0.18257
  30       x152        x15 2      0.033333       1          0.18257
  31       x161        x16 1      0.033333       1          0.18257
  32       x162        x16 2      0.033333       1          0.18257
                                                ==
                                                32
```

Relative to an optimal partial profile design with 4 of 16 attributes varying and $\boldsymbol{\beta = 0}$, the *D*-efficiency is 100%. With 120 choice sets, the raw design efficiency is 30. Relative to an optimal choice design with all attributes varying the design is 25% *D*-efficient (100 × 30 / 120 = 25). The variance matrix (not shown) is diagonal and equal to (16 / (4 × 120)) times an identity matrix. This design is optimal (assuming a main-effects partial profile model and $\boldsymbol{\beta = 0}$), but it is large. The number of choice sets is 120 = 20 blocks in the BIBD, times 18 rows in the orthogonal array, divided by 3 alternatives. You can always block a design this large. Even if you do not use it due to its size, seeing the properties of an optimal design such as this one will help you evaluate nonoptimal designs created through a computerized search.

Efficiency Directed Search for Partial-Profile Choice Designs. In this example, we will consider the same partial profile choice design problem as the preceding example. There are 16 three-level attributes, a main-effects model is fit, and the parameter vector is all zero. There are three alternatives per set with four of 16 attributes varying per set. However, this time we will use a computerized search and look for a smaller design. There are many ways to approach this problem; only one is illustrated here. The following steps make a candidate set of partial-profile alternatives with 12, 13, 14 or more ones in each row. (Since efficiency will usually increase as more values vary, some candidate alternatives are restricted to have 13 and 14 ones. Without these additional restrictions, most candidates would have 12 ones and the design would not work.) These candidates are searched to make a choice design where exactly 4 attributes vary in each set. The minimum number of choice sets with three alternatives and 16 three-level attributes is (3 – 1) × 16 / (3 – 1) = 16. In this first part of this example, we will create a partial-profile choice design with exactly 16 choice sets.

```
%macro res;
  s = (x = 1)[+];
  if       i <= 200 then bad = max(0, 12 - s);
  else if i <= 400 then bad = max(0, 13 - s);
  else                  bad = max(0, 14 - s);
  %mend;

%mktex(3 ** 16, n=600, maxtime=1, restrictions=res, seed=145, options=nosort quickr largedesign resrep)

%macro partprof;  s = 0;
  do j = 1 to 16;  s = s + all(x[,j] = 1);  end;
  bad = abs(s - 12);
  %mend;

%choiceff(data=design, model=class(x1-x16 / sta), restrictions=partprof, seed=368,
    resvars=x1-x16, nsets=16, flags=3, rscale=partial=4 of 16, beta=zero)

proc print; id set; by set; var x:; run;
```

The results are as follows:

```
                         Final Results

                         Design                    2
                         Choice Sets              16
                         Alternatives              3
                         Parameters               32
                         Maximum Parameters       32
                         D-Efficiency         2.2969
                         Relative D-Eff      57.4217
                         D-Error              0.4354
                         1 / Choice Sets      0.0625

              Variable                                         Standard
          n     Name       Label     Variance      DF         Error

          1     x11        x1 1       0.73974       1       0.86008
          2     x12        x1 2       0.34667       1       0.58879
          3     x21        x2 1       0.37003       1       0.60830
          4     x22        x2 2       0.57934       1       0.76114
          5     x31        x3 1       0.74746       1       0.86456
          6     x32        x3 2       0.71213       1       0.84388
          7     x41        x4 1       0.57606       1       0.75899
          8     x42        x4 2       0.37931       1       0.61588
          9     x51        x5 1       1.52199       1       1.23369
         10     x52        x5 2       0.45637       1       0.67555
         11     x61        x6 1       1.56920       1       1.25268
         12     x62        x6 2       1.12513       1       1.06072
         13     x71        x7 1       1.27623       1       1.12970
         14     x72        x7 2       0.55916       1       0.74777
         15     x81        x8 1       2.34167       1       1.53025
         16     x82        x8 2       1.22267       1       1.10575
         17     x91        x9 1       0.35637       1       0.59697
         18     x92        x9 2       0.55009       1       0.74168
         19     x101       x10 1      1.02839       1       1.01410
         20     x102       x10 2      0.77642       1       0.88115
         21     x111       x11 1      0.71536       1       0.84579
         22     x112       x11 2      0.45848       1       0.67711
         23     x121       x12 1      0.75706       1       0.87009
         24     x122       x12 2      0.62505       1       0.79060
         25     x131       x13 1      0.54992       1       0.74157
         26     x132       x13 2      0.49364       1       0.70260
         27     x141       x14 1      0.50734       1       0.71228
         28     x142       x14 2      0.50227       1       0.70871
         29     x151       x15 1      1.47089       1       1.21280
         30     x152       x15 2      0.89467       1       0.94587
         31     x161       x16 1      1.11194       1       1.05449
         32     x162       x16 2      0.87788       1       0.93695
                                                   ==
                                                   32
```

Relative to an optimal partial profile design with 4 of 16 attributes varying and ***β = 0***, the *D*-efficiency is 57.4217%.   One over the number of choice sets (shown in the output) is 0.0625. Multiply 0.0625 by four (since only one quarter of our attributes vary) to get a the variance to which we aspire, 4 / 16 = 0.25. Our variances are quite a bit larger. We can try again, this time asking for 24 choice sets. This is still much smaller than the 120 choice sets in the combinatorial design approach.

```
%choiceff(data=design, model=class(x1-x16 / sta), restrictions=partprof, seed=368,
   resvars=x1-x16, nsets=24, flags=3, rscale=partial=4 of 16, beta=zero)
```

The results are as follows:

```
                     Final Results

               Design                  1
               Choice Sets            24
               Alternatives            3
               Parameters             32
               Maximum Parameters     48
               D-Efficiency       4.8196
               Relative D-Eff    80.3259
               D-Error            0.2075
               1 / Choice Sets    0.0417

      Variable                               Standard
 n      Name      Label    Variance    DF      Error

 1      x11       x1 1      0.28243     1     0.53144
 2      x12       x1 2      0.27077     1     0.52036
 3      x21       x2 1      0.24277     1     0.49272
 4      x22       x2 2      0.17599     1     0.41952
 5      x31       x3 1      0.29725     1     0.54520
 6      x32       x3 2      0.19615     1     0.44289
 7      x41       x4 1      0.20679     1     0.45474
 8      x42       x4 2      0.17181     1     0.41450
 9      x51       x5 1      0.36502     1     0.60417
10      x52       x5 2      0.30437     1     0.55170
11      x61       x6 1      0.29468     1     0.54284
12      x62       x6 2      0.25777     1     0.50771
13      x71       x7 1      0.31644     1     0.56253
14      x72       x7 2      0.33579     1     0.57948
15      x81       x8 1      0.23425     1     0.48400
16      x82       x8 2      0.25974     1     0.50964
17      x91       x9 1      0.18853     1     0.43420
18      x92       x9 2      0.28211     1     0.53114
19      x101      x10 1     0.25322     1     0.50321
20      x102      x10 2     0.30579     1     0.55299
21      x111      x11 1     0.15588     1     0.39482
22      x112      x11 2     0.23657     1     0.48639
23      x121      x12 1     0.20357     1     0.45118
24      x122      x12 2     0.28216     1     0.53118
25      x131      x13 1     0.21404     1     0.46264
26      x132      x13 2     0.23343     1     0.48315
27      x141      x14 1     0.27587     1     0.52524
28      x142      x14 2     0.25788     1     0.50782
29      x151      x15 1     0.22083     1     0.46993
30      x152      x15 2     0.21324     1     0.46177
31      x161      x16 1     0.19144     1     0.43754
32      x162      x16 2     0.18856     1     0.43424
                                       ==
                                       32
```

Relative to an optimal partial profile design with 4 of 16 attributes varying and ***β = 0***, the *D*-efficiency is 80.3259%. Now our ideal variance is 4 / 24 = 0.1667. Our variances are closer to this ideal. We could get better results with more choice sets.

This is an example of a highly restricted design. The candidate set is restricted so that every candidate alternative has at least 12 ones. The choice design is restricted so that exactly 12 attributes per set have all ones. Restrictions are imposed with SAS macros that are coded to quantify the badness of each candidate row and choice set. Badness is the absolute deviation between what we have and what we want. Restrictions are very frequently used in designing choice experiments to make the subject's task easier (as in partial profiles), to make the task more realistic, and for many other reasons.

Efficiency Directed Search for Choice Designs with Attribute Overlap Restrictions. Consider a choice experiment with 3 four-level attributes, four alternatives, six choice sets, a main effects model, and $\boldsymbol{\beta = 0}$. Further assume that it is important to ensure that there is some overlap in every attribute for every choice set. Specifically, all four levels (1, 2, 3, 4) are never to appear in an attribute within choice set. Exactly three values must appear (there must be exactly one tie). Acceptable examples include (3 3 1 2), (1 1 2 3), (3 4 4 1), and so on. The following steps create and display the design.

```
%mktex(4 ** 3, n=4**3)

%macro res; do j = 1 to 3;  bad = bad + abs(ncol(unique(x[,j])) - 3); end; %mend;

%choiceff(data=design, model=class(x1-x3 / sta), restrictions=res, maxiter=10,
   resvars=x1-x3, nsets=6, flags=4, rscale=generic, beta=zero)

proc print; id set; by set; var x:; run;
```

The results are as follows:

```
                              Final Results

                        Design                  10
                        Choice Sets              6
                        Alternatives             4
                        Parameters               9
                        Maximum Parameters      18
                        D-Efficiency        4.9111
                        Relative D-Eff     81.8510
                        D-Error             0.2036
                        1 / Choice Sets     0.1667

             Variable                                          Standard
        n      Name       Label     Variance     DF       Error

        1      x11        x1 1       0.23155      1      0.48120
        2      x12        x1 2       0.21035      1      0.45864
        3      x13        x1 3       0.18117      1      0.42564
        4      x21        x2 1       0.20761      1      0.45565
        5      x22        x2 2       0.22165      1      0.47080
        6      x23        x2 3       0.19420      1      0.44068
        7      x31        x3 1       0.20703      1      0.45500
        8      x32        x3 2       0.19079      1      0.43680
        9      x33        x3 3       0.22139      1      0.47052
                                                 ==
                                                  9

                        Set    x1     x2     x3

                         1      3      1      4
                                3      1      2
                                1      2      2
                                2      3      1

                         2      1      4      3
                                2      2      3
                                4      2      1
                                1      3      2

                         3      3      4      1
                                4      4      2
                                4      3      3
```

```
         1      1      1

  4      1      2      3
         1      3      4
         2      1      2
         3      3      3

  5      2      1      3
         2      2      4
         4      4      1
         3      2      4
  6      3      1      1
         3      3      2
         2      4      4
         4      1      4
```

Relative to an optimal main-effects choice design with no restrictions and ***β = 0***, the *D*-efficiency is 81.8510%. We do not know how efficient this design is relative to the optimal restricted design. With six choice sets, the variances we aspire to in the unrestricted case are 1 / 6 = 0.1667. Our variances are larger due to the restrictions and due to the fact that we did a computerized search. Our covariances are not zero. You could create the design without restrictions to get some idea of the effect of the restrictions on efficiency. The following step searches the candidates without imposing restrictions:

```
%choiceff(data=design, model=class(x1-x3 / sta), maxiter=10, nsets=6, flags=4, rscale=generic, beta=zero)
```

The results of this step are not shown, but *D*-efficiency, assuming a main-effects model and ***β = 0***, is 96.4793. This set of restrictions is quite simple. Restrictions can be much more complicated, as we will see in the next example.

A similar design is created in this next example. However, the following restrictions are imposed:

- Attribute 1
  - 2 unique values in 2 sets (example: 1 1 2 2)
  - 3 unique values in 2 sets (example: 3 2 1 2)
  - 4 unique values in 2 sets (example: 4 2 1 3)
- Attribute 2
  - 2 unique values in 3 sets
  - 3 unique values in 2 sets
  - 4 unique values in 1 sets
- Attribute 4
  - 2 unique values in 1 sets
  - 3 unique values in 2 sets
  - 4 unique values in 3 sets

This requires that the design be restricted both within and across and choice sets. The following steps create and display the design:

```
%mktex(4 ** 3, n=4**3)

%macro res;     c = j(nsets, 3, 0);
  do i = 1 to nsets;
     if i = setnum then seti = x;    else seti = xmat[(((i - 1) # nalts + 1) : (i # nalts)),];
     do j = 1 to 3;    c[i,j] = ncol(unique(seti[,j]));    end;     end;
  bad =     abs(sum((c[,1] = 2)) - 2)  +         /* Attr 1, with 2 unique values in 2 sets */
            abs(sum((c[,1] = 3)) - 2)  +         /*         with 3 unique values in 2 sets */
```

```
          abs(sum((c[,1] = 4)) - 2)  +     /*         with 4 unique values in 2 sets */

          abs(sum((c[,2] = 2)) - 3)  +     /* Attr 2, with 2 unique values in 3 sets */
          abs(sum((c[,2] = 3)) - 2)  +     /*         with 3 unique values in 2 sets */
          abs(sum((c[,2] = 4)) - 1)  +     /*         with 4 unique values in 1 set   */

          abs(sum((c[,3] = 2)) - 1)  +     /* Attr 3, with 2 unique values in 1 set  */
          abs(sum((c[,3] = 3)) - 2)  +     /*         with 3 unique values in 2 sets */
          abs(sum((c[,3] = 4)) - 3);       /*         with 4 unique values in 3 sets */
   %mend;

%choiceff(data=design, model=class(x1-x3 / sta), restrictions=res, maxiter=10,
     resvars=x1-x3, nsets=6, flags=4, rscale=generic, beta=zero)

proc print; id set; by set; var x:; run;
```

The results are as follows:

```
                              Final Results

                        Design                    7
                        Choice Sets               6
                        Alternatives              4
                        Parameters                9
                        Maximum Parameters       18
                        D-Efficiency         4.7764
                        Relative D-Eff      79.6069
                        D-Error              0.2094
                        1 / Choice Sets      0.1667

             Variable                                              Standard
     n         Name        Label      Variance      DF              Error

     1         x11         x1 1        0.14958       1            0.38675
     2         x12         x1 2        0.26337       1            0.51320
     3         x13         x1 3        0.27383       1            0.52329
     4         x21         x2 1        0.24652       1            0.49650
     5         x22         x2 2        0.20435       1            0.45205
     6         x23         x2 3        0.23802       1            0.48787
     7         x31         x3 1        0.20328       1            0.45087
     8         x32         x3 2        0.20171       1            0.44912
     9         x33         x3 3        0.19937       1            0.44651
                                                    ==
                                                     9

                        Set    x1    x2    x3

                         1      4     1     1
                                1     3     4
                                4     4     2
                                1     1     3

                         2      1     2     1
                                4     2     3
                                4     1     4
                                1     1     2

                         3      1     2     4
                                2     3     1
                                4     2     3
                                3     3     2

                         4      2     2     4
                                1     4     3
                                3     2     1
                                1     4     1

                         5      4     1     4
                                2     3     3
                                1     2     2
                                3     4     3

                         6      3     3     4
                                2     4     4
                                4     3     2
                                2     1     2
```

Relative to an optimal main-effects choice design with no restrictions and $\boldsymbol{\beta} = \boldsymbol{0}$, the *D*-efficiency is 79.6069%. We do not know how efficient this design is relative to the optimal restricted design. With six choice sets, the variances we aspire to in the unrestricted case are 1 / 6 = 0.1667. Our variances are larger due to the restrictions and due to the fact that we did a computerized search. Our covariances are not zero.

Efficiency Directed Search for Choice Designs with Dominance Restrictions. Sometimes researchers wish to create designs while avoiding dominated alternatives. With quantitative attributes, one alternative might be dominated if every level of every attribute is less than or equal to every level of every attribute in some other alternative. We will again consider a choice experiment with 3 four-level attributes, four alternatives, six choice sets, a main effects model, and $\boldsymbol{\beta} = \boldsymbol{0}$. We will avoid dominance as follows:

```
%mktex(4 ** 3, n=4**3)

%macro res;
  do i = 1 to nalts;    do k = i + 1 to nalts;
       if all(x[i,] >= x[k,]) then bad = bad + 1;    /* alt i dominates alt k  */
       if all(x[k,] >= x[i,]) then bad = bad + 1;    /* alt k dominates alt i  */
       end;   end;
  %mend;

%choiceff(data=design, model=class(x1-x3 / sta), restrictions=res, maxiter=10,
        seed=495, resvars=x1-x3, nsets=6, flags=4, rscale=generic, beta=zero)

proc print; id set; by set; var x:; run;
```

The results are as follows:

```
                              Final Results

                        Design                   7
                        Choice Sets              6
                        Alternatives             4
                        Parameters               9
                        Maximum Parameters      18
                        D-Efficiency        5.3831
                        Relative D-Eff     89.7189
                        D-Error             0.1858
                        1 / Choice Sets     0.1667

               Variable                                       Standard
          n      Name       Label     Variance     DF           Error

          1      x11        x1 1       0.29221      1         0.54056
          2      x12        x1 2       0.18099      1         0.42543
          3      x13        x1 3       0.17525      1         0.41863
          4      x21        x2 1       0.20986      1         0.45810
          5      x22        x2 2       0.18696      1         0.43238
          6      x23        x2 3       0.20933      1         0.45753
          7      x31        x3 1       0.23567      1         0.48546
          8      x32        x3 2       0.17876      1         0.42280
          9      x33        x3 3       0.19543      1         0.44207
                                                   ==
                                                    9

                        Set    x1     x2     x3

                         1      2      3      3
                                1      4      1
                                4      1      2
                                3      2      4

                         2      4      1      2
                                3      1      3
                                2      2      1
                                1      4      4

                         3      1      2      3
```

```
           3     4     2
           4     3     1
           2     1     4

   4       2     4     2
           4     2     1
           3     3     2
           1     1     3

   5       1     3     4
           1     4     3
           3     1     1
           2     2     2

   6       3     4     1
           4     2     3
           2     1     4
           1     3     2
```

Relative to an optimal main-effects choice design with no restrictions and $\boldsymbol{\beta = 0}$, the *D*-efficiency is 89.7189%. Our variances are larger than 1/6 due to the restrictions and due to the fact that we did a computerized search. Our covariances are not zero.

Constant Alternative. In this example, the design from the preceding example is modified to include a "none of these" constant alternative. The following steps create the design:

```
%mktex(4 ** 3, n=4**3)

data cand;
  retain f1-f3 0 f4 1;
  if _n_ = 1 then do; output; f4 = 0; f1 = 1; f2 = 1; f3 = 1; end;
  set design;
  output;
  run;

proc print data=cand(obs=5); run;

%macro res;
  do i = 1 to nalts - 1;    do k = i + 1 to nalts - 1;
      if all(x[i,] >= x[k,]) then bad = bad + 1;    /* alt i dominates alt k  */
      if all(x[k,] >= x[i,]) then bad = bad + 1;    /* alt k dominates alt i  */
      end;   end;
  %mend;

%choiceff(data=cand, model=class(x1-x3 / zero=none), restrictions=res, maxiter=10,
       seed=121, resvars=x1-x3, nsets=6, flags=f1-f4, beta=zero)

proc print; id set; by set; var x:; run;
```

The DATA step creates a candidate set with a constant alternative followed by the 64 candidates that vary. The first five candidates are as follows:

```
Obs    f1    f2    f3    f4    x1    x2    x3

  1     0     0     0     1     .     .     .
  2     1     1     1     0     1     1     1
  3     1     1     1     0     1     1     2
  4     1     1     1     0     1     1     3
  5     1     1     1     0     1     1     4
```

Missing values are used for the constant alternative. Other levels can be assigned later. In the flag variables that flag the alternatives for which each candidate can be used, f4 = 1 for the constant alternative only, and f1 = f2 = f3 = 1 while f4 = 0 for all of the other candidates. The rest of the steps are similar to the preceding example. The ZERO=NONE coding option is used so that a coded variable is created for each nonmissing level. This is a good way to ensure that you know precisely which parameters are estimable. The variances are as follows:

```
      Variable                                        Standard
 n      Name      Label    Variance     DF       Error

 1      x11       x1 1      2.66315      1      1.63192
 2      x12       x1 2      3.50756      1      1.87285
 3      x13       x1 3      4.04671      1      2.01164
 4      x14       x1 4      4.09053      1      2.02251
 5      x21       x2 1      2.42013      1      1.55568
 6      x22       x2 2      2.22238      1      1.49077
 7      x23       x2 3      2.03376      1      1.42610
 8      x24       x2 4       .           0       .
 9      x31       x3 1      2.35369      1      1.53417
10      x32       x3 2      2.18059      1      1.47668
11      x33       x3 3      2.49307      1      1.57895
12      x34       x3 4       .           0       .
                                        ==
                                        10
```

There is one additional parameter that can be estimated due to the constant alternative. You can specify the option “DROP=x24 x34” to drop the terms that cannot be estimated. See Kuhfeld (2010a) for more about constant alternatives and coding. The first choice set is as follows:

```
Set    x1    x2    x3

 1      3     1     3
        1     3     2
        2     4     1
        .     .     .
```

Combinatorial Approach to MaxDiff Designs. In a MaxDiff experiment, subjects see *t* alternatives, shown in *b* sets of size *k*. In each set, they choose the best and worst alternative. A BIBD or an efficient design that does not quite have a perfect BIBD structure can be used. The following step creates a design for *t*=12 alternatives with *b*=22 sets of size *k*=6.

```
%mktbibd(nattrs=12, nsets=22, setsize=6, seed=522)
```

The results are as follows:

```
Block Design Efficiency Criterion       100.0000
Number of Attributes, t                       12
Set Size, k                                    6
Number of Sets, b                             22
Attribute Frequency                           11
Pairwise Frequency                             5
Total Sample Size                            132
Positional Frequencies Optimized?            Yes

          Attribute by Attribute Frequencies

           1  2  3  4  5  6  7  8  9 10 11 12

        1 11  5  5  5  5  5  5  5  5  5  5  5
        2    11  5  5  5  5  5  5  5  5  5  5
        3       11  5  5  5  5  5  5  5  5  5
        4          11  5  5  5  5  5  5  5  5
        5             11  5  5  5  5  5  5  5
        6                11  5  5  5  5  5  5
        7                   11  5  5  5  5  5
        8                      11  5  5  5  5
        9                         11  5  5  5
       10                            11  5  5
       11                               11  5
       12                                  11
```

```
Attribute by Position Frequencies

            1  2  3  4  5  6

        1   2  2  2  1  2  2
        2   2  2  2  2  1  2
        3   2  1  2  2  2  2
        4   2  2  2  2  2  1
        5   2  2  2  2  1  2
        6   2  2  2  1  2  2
        7   2  1  2  2  2  2
        8   1  2  2  2  2  2
        9   2  2  2  2  2  1
       10   2  2  1  2  2  2
       11   2  2  1  2  2  2
       12   1  2  2  2  2  2

Balanced Incomplete Block Design

x1    x2    x3    x4    x5    x6

 7    12     8    11     3     2
 9     4     2    10     6     8
 1     8     6     4    11    12
11     7     4     9     3     8
 3     6     1     8    12     2
 4     2     8     5    10     3
 2     9     6     4     7     1
 9     3     2     5     1    11
10     6     5     8     1     7
 8     1     7    11     4     5
 4     1     3     7    12    10
10     9    12     7     8     5
 7    10    11     3     9     6
 5    12     9     3     8     6
 2    11    10    12     5     4
 1     5    12     2     7     9
 5    10     1     6    11     3
 3     5     4     2     6     7
11     8     9    10     2     1
 6    11     5     9     4    12
 6     2     7    12    10    11
12     4     3     1     9    10
```

The block design efficiency criterion is 100, so this is a BIBD. Each alternative occurs equally often (11 times), and each alternative occurs with every other alternative five times. If we were to try again with *b*=20 blocks instead of *b*=22 (not shown), we get an efficiency criterion of 99.9528 percent, alternative frequencies of 10, and alternative, by alternative frequencies of 4 and 5. A BIBD is not possible with this specification. Still, this design is probably fine for practical use.

Coding and Efficiency. Every example in this paper uses a standardized orthogonal contrast coding of the effects. This is a particularly convenient coding to use for design efficiency evaluation, because in many cases, it is easy to know the maximum *D*-efficiency for a design that is coded this way. However, most researchers do not use this coding for analysis. Instead, they prefer codings such as reference cell or effects coding. These codings do not provide an orthogonal decomposition of each effect even with an orthogonal design, so they are less convenient for design evaluation. Different coding schemes change the raw but not the relative *D*-efficiency of competing designs. Consider an optimal design created by combinatorial means. Evaluate the design using four different codings and call the raw *D*-efficiency values *a*, *b*, *c*, and *d*. Now create a competing design by some other method and evaluate it with the same four coding schemes. Call the raw *D*-efficiency values *p*, *q*, *r*, and *s*. Now compare the relative *D*-efficiencies for the two designs under the four different coding schemes. You will find that *p/a* =

q/b = *r/c* = *s/d*, which equals the relative *D*-efficiency (scaled to a proportion) found by using the standardized orthogonal contrast coding. While it was not shown in any examples, it is good practice to evaluate the design using the actual coding that will be used in the analysis. The same number of parameters will be estimable, but you will see information about the actual variances (given your assumed parameter vector). These variances can be further scaled by specifying the anticipated number of subjects (using the **n=** option in the ChoicEff macro). This scaling changes all variances by the same amount.

Summary and Conclusions

This paper presents a number of examples of different types of choice designs (main-effects, alternative, specific, partial profile, and MaxDiff) and different construction methods. Approaches include combinatorial methods (which make 100% *D*-efficient designs, but come in limited sizes and are often unrealistic), efficiency directed searches (designs that do not have size limitations, but are typically less than 100% *D*-efficient), and restricted designs (where you can impose restrictions to make the design more realistic or easier for the subject, while paying a price in *D*-efficiency). There is no uniformly superior method. The tool that you will use will depend on your problem. For some problems, we can measure *D*-efficiency on a known 0 to 100 scale. Other times we cannot. Part of the art of choice design is understanding how to interpret the variances and other information to evaluate the goodness of your design. No matter how you construct your design, it is important that you evaluate it to ensure that it has the properties that you will need for analysis (estimable parameters with reasonable variances).

References

Anderson, D.A. (2003), personal communication.

Chrzan, K. and Elrod, T. (1995) “Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes,” paper presented at the AMA's 1995 Advanced Research Techniques Forum, Monterey, CA.

Chrzan, K.,   Zepp, J., and White, J. (2010) “The Success of Choice-Based Conjoint Designs Among Respondents Making Lexicographic Choices”, Proceedings of the Sawtooth Software Conference, http://www.sawtoothsoftware.com/download/techpap/2010Proceedings.pdf, visited 2/21/2012.

Finn, A., and Louviere, J., (1992), “Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety,” *Journal of Public Policy & Marketing*, Vol. 11, 1, 12-25.

Gilbride, T. and Allenby, G., (2004), “A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules,” *Marketing Science*, 23, 3, 391-406.

Hauser, J., Dahan, E., Yee, M., and Orlin, J., (2006), “Must Have’ Aspects vs. Tradeoff Aspects in Models of Customer Decisions,” *Sawtooth Software Conference Proceedings,* 169-181.

Huber, J. and Zwerina, K. (1996), “The Importance of Utility Balance in Efficient Choice Designs,” *Journal of Marketing Research*, 33, 307–317.

Kuhfeld, W.F. (2010a) Marketing Research Methods in SAS, http://support.sas.com/resources/papers/tnote/tnote_marketresearch.html, visited 2/21/2012.

Kuhfeld, W.F. (2010b) Orthogonal Arrays, http://support.sas.com/techsup/technote/ts723.html, visited 2/21/2012.

Kuhfeld, W., Tobias, R., and Garratt, M. (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, 31, 545–557.

Lazari, A., Anderson, D., (1994), "Design of Discrete Choice Set Experiments for Estimating Both Attribute and Availability Cross Effects," *Journal of Marketing Research*, Vol. 31, (August), 375-383.

Louviere, J., and Woodworth, G., (1983), "Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data," *Journal of Marketing Research*, 20, 350-367.

Louviere, J. J. (1991), "Best-Worst Scaling: A Model for the Largest Difference Judgments," Working Paper, University of Alberta.

Street, D.J., and Burgess, L. (2007) The Construction of Optimal Stated Choice Experiments, New York: Wiley.

IN DEFENSE OF IMPERFECT EXPERIMENTAL DESIGNS: STATISTICAL EFFICIENCY AND MEASUREMENT ERROR IN CHOICE-FORMAT CONJOINT ANALYSIS

F. REED JOHNSON,
JUI-CHEN YANG, AND
ATEESHA F. MOHAMED
RESEARCH TRIANGLE INSTITUTE

ABSTRACT

Efficient experimental designs ensure that researchers can maximize the precision of choice-model estimates for a given sample size. In many cases, however, response or measurement error may be a greater source of estimation imprecision than statistical error. This study reports the results of simulating effects of sample size, experimental-design features, and other study characteristics on the precision of choice-model estimates using re-sampled data from 29 conjoint studies. We conclude that more efficient experimental designs contribute relatively little to precision of choice-model estimates, suggesting that researchers should pay more attention to reducing choice-task complexity and devote more resources to larger sample sizes.

INTRODUCTION

For any choice-format conjoint analysis (CA) study, researchers must make a number of decisions about which attributes, how many levels, how many alternatives per choice task, how many choice tasks, and which choice-task format to use. Unfortunately there is little evidence on the effects of these factors on data quality to guide study-design decisions. Several studies have shown that task complexity often leads to inconsistency and higher variance (DeShazo and Fermo, 2002; Johnson et al., 1998, 2000; Özdemir et al., 2010; Saelensminde, 2002; Swait and Adamowicz, 2001). More attributes provide more trade-off information, but also increase complexity and encourage simplifying heuristics. More levels for each attribute provide more preference granularity, but larger sample sizes may be required to estimate the additional parameters. More alternatives per choice task also increase task complexity. More choice questions yield more data for a given sample size, but can contribute to respondent fatigue, with associated degradation of data quality.

Because overall precision depends on both response error and statistical error, larger sample sizes can help compensate for response error. The intent of our study was to quantify the relative importance of statistical efficiency related to sample size and experimental-design features (called the "vertical dimension" by Huber in this volume) after controlling for study characteristics. We employ a strategy of simulating different sample sizes for 29 actual conjoint studies of patient preferences for health interventions.

STATISTICAL EFFICIENCY, EXPERIMENTAL DESIGNS, AND SAMPLE SIZE

Greater statistical efficiency yields smaller variance in parameter estimates in choice models. Improved statistical efficiency with a given sample size can be obtained with more efficient experimental designs. In general, efficient designs are orthogonal, are balanced in level counts, account for correct priors on parameter values, and account for intended functional-form specifications and estimation approach (Kuhfeld, 2010; Kanninen, 2002; Rose et al., 2008). Avoiding overlapped attribute levels and dominated alternatives also improve design efficiency.

While better experimental designs improve statistical efficiency for a given sample size, research sponsors often ask "How large does the sample need to be?" presumably given an efficient experimental design. The answer rarely is straightforward because there are no power calculations for choice models such as those for effect-size calculations in simple experiments.[23] Sample-size considerations are likely to be influenced not only by desired precision, desired representativeness, and need for subsample analysis, but also by budget considerations. A more appropriate question thus is "What is the right sample size?".

Most discussions about the right sample size focus on statistical error. However, there are other sources of error that may be even more important in determining estimate precision. In addition to statistical error, precision is affected by modeling error, sampling error, and response (or measurement) error. Statistical error varies inversely with the square root of the sample size. Modeling error results from biased estimates in poorly specified models. Also, as in survey research generally, nonrandom sampling procedures can result in biased, although not necessarily imprecise, estimates. Response (or measurement) error can arise from poor attribute definitions, choice-task complexity, and respondents' inattention and fatigue. Thus minimizing measurement error, or maximizing response efficiency, relates directly to the unique study-design decisions required by conjoint studies.

RESPONSE EFFICIENCY-DESIGN TRADEOFFS

Response efficiency depends on how well survey respondents interpret and respond to choice questions (called the horizontal dimension or "difficulty of the choice to the respondent" by Huber in this volume). Task difficulty typically is determined by the number and complexity of attributes, the number of alternatives per choice task, and the number of choice tasks. The more difficult choice tasks are, the greater the response (or measurement) error. In general, task complexity can be mitigated by (a) reducing the number of tasks per survey respondent, which could require a compensating larger sample size, (b) reducing complexity with attribute overlaps which reduces statistical efficiency, and/or (c) employing nested tasks (e.g., first show a choice between A and B followed by a choice between A/B and C), which increases analytical difficulty.

While choice difficulty obviously depends on how many attributes and alternatives subjects must evaluate simultaneously, implausible attribute-level combinations also are a common source of response inefficiency. Choice questions that are not self-explanatory or do not look realistic also increase choice difficulty. For example, suppose there are two attributes of interest – pain experienced and restrictions on activities of daily living. Each attribute has three levels. For pain experienced, the three levels are mild, moderate, and severe pain. For restrictions on activities of daily living, the three levels are mild, moderate, and severe restrictions. Table 1

[23] However, Bliemer and Rose (2005) have derived a measure called S-efficiency that optimizes experimental designs for given sample sizes.

presents possible combinations of the attribute levels. Profile #3 combines mild pain with severe restrictions on activities of daily living. Profile #7 combines mild restrictions on activities of daily living with severe pain. Based on reasonable definitions of pain severity and daily-activity restrictions, both of these combinations would appear implausible to most survey respondents. Survey respondents could react to this implausibility in several ways, such as refusing to answer the question, "recoding" the levels by replacing one of the conflicting levels with a more plausible label, or ignoring one of the attributes.

While including Profiles #3 and #7 ensures orthogonality of the experimental design and thus improves design efficiency, orthogonality comes at the expense of a behavioral response that increases the variance and possibly even ability to estimate the parameters of the choice model. Following Huber's (in this volume) recommendation to "make sure your design parameters generate choices that respondents can handle," we might consider excluding Profiles #3 and #7 from the design set. However, removing these implausible combinations creates correlations between attribute levels. In this example, the correlation is 0 when all combinations of attribute levels are allowed. Removing Profiles #3 and #7 from the design set induces a correlation of 0.5 (see Table 2 for details).

Table 1. Possible Combinations of Pain Severity and Restrictions on Activities of Daily Living

Profile	Pain Experienced	Restrictions on Activities of Daily Living
1	Mild	Mild
2	Mild	Moderate
3	Mild	Severe
4	Moderate	Mild
5	Moderate	Moderate
6	Moderate	Severe
7	Severe	Mild
8	Severe	Moderate
9	Severe	Severe

Table 2. Effects of Constraining Design

Number of Omitted Combinations	Correlation
0	0.00
1	0.23
2	0.50
3	0.59
4	0.85
5	1.00

Sample-Size Rules of Thumb

Power calculations are complex for CA studies because precision depends on the factors discussed above. CA researchers often apply rules of thumb or past experience to determine the appropriate sample size. For example, Hensher et al. (2005) recommend that a minimum of 50 respondents choose each alternative in a labeled design. For a 2-alternative unlabeled design with choice probabilities set to 75% for one alternative and 25% for the other alternative (Kanninen, 2002), the Hensher rule results in a minimum sample size of 200.

Another rule of thumb is based on results from several marketing studies (Johnson and Orme, 1996). Orme (2010) subsequently reported Johnson's suggestion that the number of observations available to estimate each parameter should be ≥ 500 so that:

$$\frac{n \times t \times a}{c} \geq 500$$

where n is the number of respondents, t is the number of choice tasks, a is the number of alternatives, and c is the largest number of levels in any attribute. Table 3 presents the minimum sample required for 2-alternative tasks when maximum number of levels and number of choice tasks vary. For a study with 2 alternatives, with 4 being the largest number of levels in any attribute and with 8 choice tasks, this rule indicates a minimum sample size of 125.

Table 3. Sample Sizes for a Two-Alternative Task Using Johnson's Rule of Thumb

Number of Choice Alternatives = **2**

Number of task repetitions =	5	6	7	8	9	10	11	12
Max number of levels for any one attribute =	Number of Respondents >=							
2	100	83	71	63	56	50	45	42
3	150	125	107	94	83	75	68	63
4	200	167	143	125	111	100	91	83
5	250	208	179	156	139	125	114	104
6	300	250	214	188	167	150	136	125

Such rules of thumb may provide a useful general guide, but still rely on researchers' experience to judge the effect of specific study-design features on the likely precision of choice-model estimates. Ideally, one would pool data from a large number of studies and perform a statistical meta analysis of the effect of study characteristics on estimated precision. However, the number observations that would be required to control for the large heterogeneity among studies would be daunting. In any event, we are unaware of any such database. Instead, after first summarizing findings of a recent review of published studies in health, we describe a simulation analysis based on data from 29 conjoint applications in health and healthcare undertaken by researchers at Research Triangle Institute (RTI).

Study Designs in Health Applications

Marshall et al. (2010) reviewed all English-language CA studies in health and healthcare published between January 2005 and July 2008. A total of 78 CA studies were reviewed. These studies had sample sizes ranging between 13 and 1258. The mean sample size was 259, with a mode of 119. In terms of study-design features (number of attributes, maximum number of levels, and number of choice tasks), there was some similarity across the 78 studies as indicated in Table 4 below.

Table 4. Summary of Survey Design Characteristics (N = 78)

Statistic	Number of Attributes	Maximum Number of Levels	Number of Choice Tasks
Mean	6	4	12
Median	6	4	10
Mode	6	3	8
Minimum	3	2	3
Maximum	16	6	36

The RTI Studies

The meta-dataset consisted of 29 CA surveys with a similar "look and feel" that generated 31 datasets. These studies spanned several therapeutic areas, including Alzheimer's disease, asthma, bipolar disorder, Crohn's disease, epilepsy, gastrointestinal stromal tumor, hemophilia, hepatitis B, hepatitis C, idiopathic thrombocytopenic purpura, irritable bowel syndrome, kidney transplant, migraine, multiple sclerosis, obesity, osteoarthritis, psoriasis, rheumatoid arthritis, rotavirus, type 2 diabetes mellitus, colorectal cancer, HIV, and renal cell carcinoma. The studies also varied in instrument complexity, including difficult concepts such as probabilistic outcomes in several studies. Inevitably, some features were confounded. For example, samples in cancer studies consist exclusively or predominantly of older respondents.

In general, the studies included five to nine attributes with the number of levels varying between two and four. Almost all the studies (27/29) implemented two-alternative tasks, and the number of choice tasks varied between 8 and 13. Survey respondents were from the United States, Canada, and Europe. All surveys were administered online.

Sample-Size Simulations

The design for the meta-analysis was inspired by a Bryan Orme suggestion: "If you find yourself questioning whether you really needed to have such a large sample size, delete a random subset of the data to see how fewer respondents would have affected the findings" (Orme, 2010). We generalized Orme's idea by drawing random samples of different sizes with replacement from multiple datasets. Figure 1 presents the steps in our sample-size simulations. For each of the 31 datasets, we simulated 15 sample sizes ranging from 25 to 1000 by sampling with replacement from the original dataset. For each of 10,000 draws, we estimated a main-effects conditional-logit model and calculated the difference between the best and worst profile utility. For each sample size, we calculated the mean and standard deviation of the 10,000 simulated utility differences.

Figure 1. Steps in Sample-Size Simulations

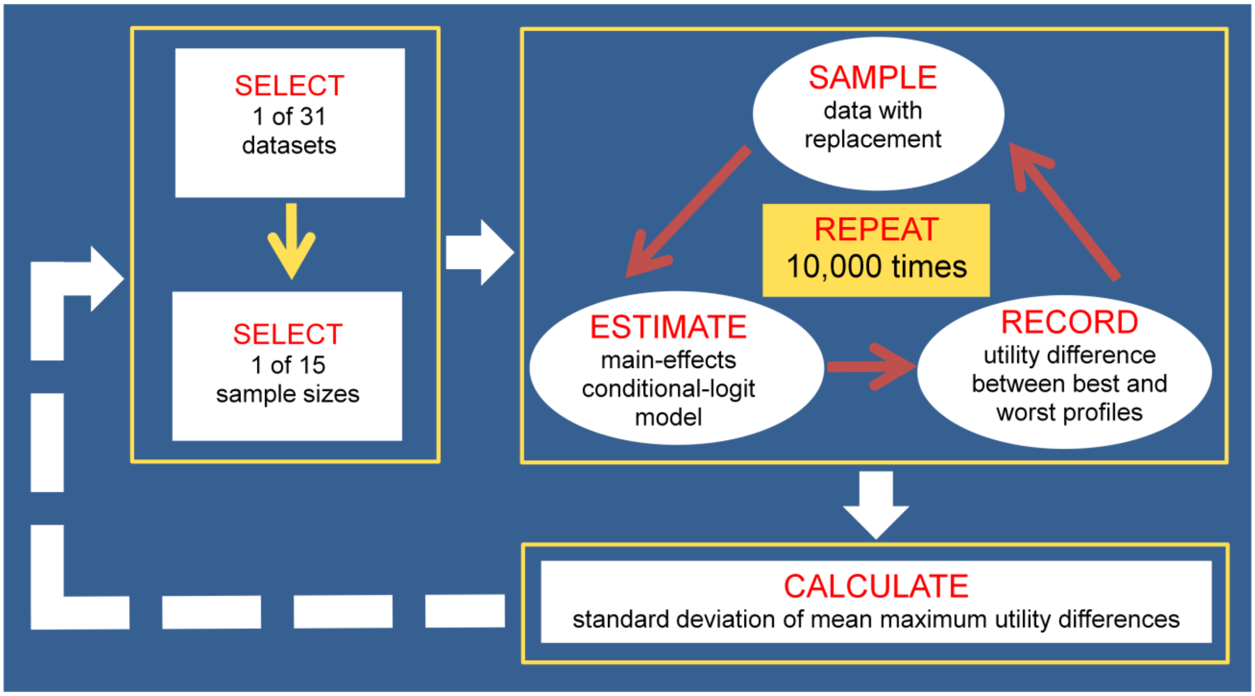
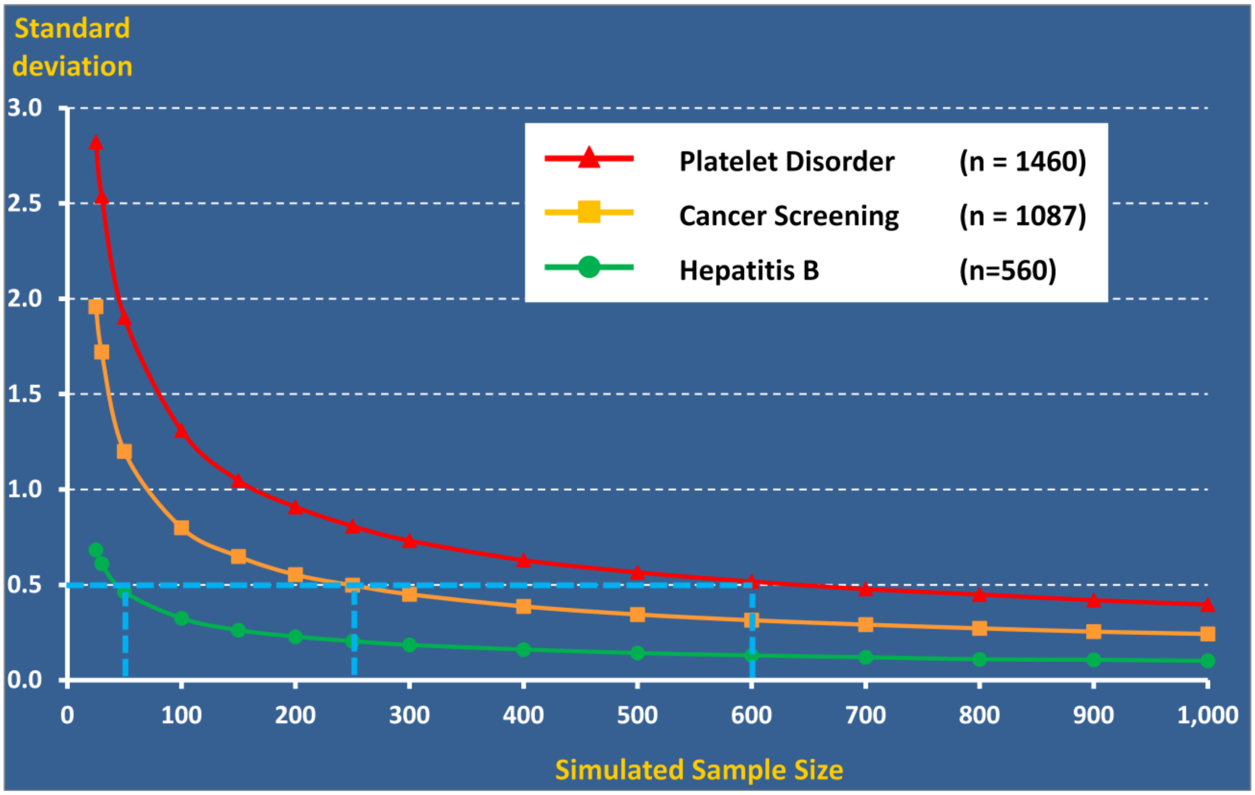


Figure 2 shows example plots of the simulated standard deviations, our normalized measure of precision, against simulated sample sizes for three studies. The plots confirm the success of the simulation. Normalized choice-model precision is inversely related to the square root of the simulated sample size (McFadden, 1974). Thus when sample sizes are small (<100), increases in sample size result in large improvements in precision. However, when sample sizes are large (>500) increases in sample size provide only minimal improvements in precision.

While all three plots have similar shapes with respect to sample size, their horizontal placement varies. This placement reflects differences in precision from study-specific factors other than sample size. For example, to achieve a normalized precision score of 0.5, the platelet-disorder survey would require a sample size of about 650. In contrast, the hepatitis B survey would require a sample size of only about 50. Subsequent analysis of the data provided insights about what factors influence precision after controlling for the effects of sample size.

Figure 2. Simulated Results for 3 Studies



META-REGRESSION RESULTS

We estimated an ordinary least-squares model to examine to what extent sample size and study-design features affect precision. The dependent variable was the simulated standard deviation for each study and simulated sample size. Independent variables included simulated sample sizes, measures of experimental-design efficiency and study-design features, and study characteristics.

Table 5 summarizes the mean values for each independent variable, the significance of the independent variables in the meta-regression, as well as the change in precision for a given change in an independent variable. For example, assuming a sample size of 250, precision improved by 20% by increasing the sample size by 100. All explanatory variables were modeled as linear, except for D-score and number of choice tasks, which had significant squared terms. More choice tasks increased precision up to a point but too many choice tasks reduced precision. The results indicated that 9 or 10 is the optimal number of choice tasks per respondent (inflection point = 9.5). D-score, most number of levels for any attribute, inclusion of an opt-out or no-purchase alternative, and a dummy for whether the sample was recruited from a patient-support organization had significant interactions with sample size.

Table 5. Meta-Regression Estimates

Variable	Mean	Change	Precision Effect
1/sqrt(N)***	(250)	–100	-35.0%
		+100	+19.7%
N x D-Score*** **N x D-Score Squared***	21	+25%	+2.3%
Number of Attributes***	6.3	+1	-5.6%
Number of Probabilistic Attributes***	1.8	+1	-7.5%
N x Most Number of Levels<4***	0.02	4 to less than 4	-18.5%
N x Most Number of Levels=4***	.43		
N x Most Number of Levels>4***	.56	4 to greater than 4	+19.0%
Number of Choice Tasks** **Number of Choice Tasks Squared****	9.2	6 to 9	+9.3%
		9 to 12	-5.5%
N x Opt-out Alternative***	0.13	Yes	+12.2%
N x Patient-Group Sample***	0.29	Yes	-9.4%

Note: *** p < 0.01, ** p < 0.05, * p < 0.10

EXPERIMENTAL DESIGN AND PRECISION

Several study-design features had statistically significant impacts on model precision. Such features presumably affect both task complexity and design efficiency. Controlling for study-design features such as number of attributes, levels, and choice tasks, a 25% increase in D-score relative to the mean improved precision by only 2.3%. Table 6 shows that the model explains 84% of the variance in the simulated data. Experimental-design efficiency accounts for only 8% of the explained variance in the model. Study-design and other features account for more than twice as much of the variance as design efficiency. However, sample size is by far the most important factor in explaining model precision, accounting for more than 7 times as much of the variance as design efficiency.

Table 6. Explained Variance in Meta Regression

Type of Variable	Explained Variance	Cumulative
Experimental-Design Efficiency (D-score)	8%	15%
Study-Design Features (Number of tasks, most levels, number of attributes, opt-out)	7%	7%
Other Features (Probabilistic attributes, disease dummies, patient group sample)	11%	26%
Sample size	58%	84%

DISCUSSION

Our study has three main findings. First, there appears to be a rough consensus among CA researchers in health on the workable ranges for number of attributes, the number of levels for each attribute, and the number of choice tasks. This consensus presumably reflects experience with numerous trials that test feasibility of various study designs. Second, conventional wisdom suggests that experimental design is very important. There is a good case for near-orthogonality to ensure parameters of interest in the choice model are estimable, and more efficient designs logically improve statistical power. However, while more complex designs may improve design efficiency, such gains could be offset by decreases in response efficiency. The meta-analysis of RTI studies provides weak support for the importance of design efficiency after controlling for study features that affect task complexity. Third, these results imply that researchers should reconsider the allocation of resources in survey development and data collection. Studies that use more tractable choice tasks and somewhat larger samples are likely to yield more precise choice-model estimates than studies with highly efficient experimental designs, but more complex choice tasks and somewhat smaller sample sizes.

These conclusions are subject to some obvious caveats. The simulation experiment based on RTI studies has the advantage of a degree of uniformity in the look and feel of the instruments used to generate the meta-data. This advantage also is the greatest disadvantage of the study. The ranges of study variables are rather narrow and do not reflect the wider ranges of study features found in published studies. Furthermore, the experimental designs in all RTI studies were based on near-orthogonal, D-efficient designs. Our results do not provide insights into the possible effects of randomized designs or other approaches that have been developed over the last few years.

References

Bliemer MCJ, Rose JM. Efficiency and sample size requirements for stated choice studies. 2005. Available from: http://ws.econ.usyd.edu.au/itls/wp-archive/itls_wp_05-08.pdf. [Accessed April 25, 2012].

DeShazo JR, Fermo G. 2002. Designing choice sets for stated preference methods: the effects of complexity on choice consistency. Journal of Environmental Economics and Management 44: 123–143.

Hensher DA, Rose JM, Greene WH. *Applied Choice Analysis*. Cambridge: Cambridge University Press; 2005.

Johnson FR, Banzhaf MR, Desvousges WH. 2000. Willingness to pay for improved respiratory and cardiovascular health: a multiple-format, stated-preference approach. Health Economics 9: 295–317.

Johnson FR et al. 1998. Eliciting stated health preferences: an application to willingness to pay for longevity. Medical Decision Making 18: S57–67.

Johnson R, Orme BK. 1996. How many questions should you ask in choice-based conjoint studies? Presented at the American Marketing Association 1996 Advanced Research Techniques Forum.

Kanninen B. Optimal design for multinomial choice experiments. J Mark Res 2002; 39:214–227.

Kuhfeld WF. Experimental design: efficiency, coding, and choice designs. 2010. Available from: http://support.sas.com/techsup/technote/mr2010c.pdf. [Accessed April 25, 2012].

Marshall D, Bridges J, Hauber AB, Cameron R, Donnally L, Fyie, K, Johnson FR. 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; 3:249-256.

McFaddden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Frontiers in Econometrics.New York: Academic Press; 1974. p. 105-42.

Orme BK. Getting started with conjoint analysis: strategies for product design and pricing research. 2nd ed. Madison, WI: Research Publishers LLC; 2010.

Özdemir S, Mohamed AF, Johnson FR, Hauber AB. 2010. Who pays attention in stated-choice surveys? Health Economics 19(1):111-118.

Rose JM, Bliemer MCJ, Hensher DA, et al. Designing efficient stated choice experiments in the presence of reference alternatives. Transp Res Part B Methodol 2008 May;42(4):395-406.

Saelensminde K 2002. The impact of choice inconsistency in stated choice studies. Environmental and Resource Economics 23: 403–420.

Swait J. and Adamowicz W. 2001. Choice environment, market complexity, and consumer behavior: A theoretical and empirical approach for incorporating decision complexity into models of consumer choice. Organizational Behavior and Human Decision Processes 86(2):141–167.

CBC DESIGN FOR PRACTITIONERS: WHAT MATTERS MOST

JOEL HUBER
DUKE UNIVERSITY

Choice design has generated a great deal of heat, with strong advocates for particular ways to set up a choice design. I focus on the characteristics of the design that the respondent sees—how many alternatives per choice, how many attributes differing per choice, and how many choices per respondent. My central theme will be that these need to be determined by consideration of how difficult the choice task is to the respondent. The more novel, emotional or complex the task is for the respondent, the greater the need to provide respondent training and reduce the number of tasks, the number of alternatives per task and the number of attributes that differ per alternative. In such a process there is a critical role for optimization routines that minimize D-error given a set of design characteristics, but they largely serve to provide a balanced design and to make sure that critical interactions or nonlinearities can be estimated. To show the importance of choice difficulty, I will give examples of contexts in which respondents were only able to cope with choice pairs that differ on two attributes and extreme times when it makes sense to ask only one choice.

The paper will be organized the following way. First, I explore two divides that separate conjoint users—the first related to statistically optimal vs. respondent feasible designs and the second to whether the choice attributes are novel and emotionally difficult vs. familiar and easy to trade off.

DIFFERENT DESIGN NEEDS AMONG CHOICE-BASED CONJOINT USERS

The divides can be characterized by a horizontal and a vertical dimension, where the horizontal dimension relates to the difficulty of the choice to the respondent and the vertical dimension relates to the statistical vs. practical efficiency of the design. Both of these dimensions reflect different orientations, resources and skills among its proponents, and lead to quite different designs.

Consider first the horizontal dimension, reflected in the heterogeneity in the audience at a Sawtooth Software conference. On one side are those who use conjoint to estimate standard marketing parameters, such as the price or shelf elasticity for a product that they sell. Often the product is in stores and consumers already know about the brand features and have pretty clear expectations about prices. There is little need to get a person thinking about the attributes. Further, for categories with wide distribution like cat litter, detergent or fast food it is not hard to interview respondents who are in the market. Accordingly, for those on that side of the divide, questions can be few, familiar and fast. To the extent that the parameters involve aggregate statistics, such as cross price elasticities or cannibalization effects, data at the aggregate level from relatively few tasks per respondent may be sufficient.

Somewhere in the middle along this horizontal divide are conjoint studies where most of the attributes are known but a few focal ones are novel and require effort on the part of respondents to assess how they would trade them off. Consider the case of new product development, where it is important to assess the impact of a new feature or service. Such studies can require elaborate preparation, as where a motorcycle manufacturer desires to test a novel handling ability or extreme styling. Conjoint exercises may be implemented in clinics where respondents have the ability to compare the look and feel of such innovations. In such cases each respondent is typically well compensated so they are willing to put up with more demanding conjoint tasks, and fill out lengthy questionnaires about purchases, media use and demographics.

At the other side of the horizontal divide reside conjoint studies that are profoundly difficult for respondents. These are represented in the health choice work done by Saigal and Dahan and the work done by Johnson, Yang and Mohamed (both in this volume). Beyond health care, similar concerns arise in conjoint studies of life options, dating preferences, environmental changes, and financial decisions. Such decisions can be difficult in three ways. First, they may simply be novel, as with the implications of global positioning capabilities for cell phones. Second, and more problematic, the decision may evoke strong and negative emotions, as in the choice among surgical procedures or insurance options. Finally, the decisions may require trading off options that are fundamentally difficult to resolve. Consider attributes that are expressed as probabilities against those that involve outcomes at a later and uncertain time. How should one decide between a 5% decrease in the probability of death this year against the certainty of being employed? Neither expertise nor long thought are much use against this kind of problem.

For difficult choices, effort is necessary to help respondents prepare for the choice and understand how they feel about the attributes. Further, choices must be simple or respondents may answer randomly or refuse to participate. Simplicity in this case may lead to choices between pairs and only a few attributes, and in some case may justify only one choice per respondent. I will give examples of cases where such simplicity is warranted below.

Summarizing, the horizontal divide separates those studies whose conjoint problems are relatively easy for respondents from those that are difficult. The latter require substantial efforts to prepare respondents for the choices. Additionally, the choices need to be simpler in the sense of having few alternatives or attributes. However, deep questioning and high cost per respondent justify getting substantial information per respondent By contrast, those using choice-based conjoint to study minor changes in package goods or familiar services can and should expose respondents to choices that mirror the complexity of multiple attributes available in the marketplace. As indicated by Kurz and Binner (this volume) such studies are often conducted with relatively low cost web panels whose members are reluctant to complete long questionnaires or more that 4-6 choice tasks. Such short instruments may not be such a problem if the goal is to assess aggregate rather than individual behavior. The Bayesian estimates will appropriately compensate for relatively poor resolution at the individual level.

Consider next the vertical dimension, anchored by the brain at the top and feet (of clay) at the bottom. This divide is less across the types of business problems and more about the elegance vs. serviceability of the choice design. It is easy to become enamored of braininess of beautiful experimental designs. I still recall the room in which I was sitting in a library at Penn when I saw my first orthogonal array. Knowing they are optimal brings even more satisfaction. Kuhfeld and Wurst (this volume) present the many ways that the SAS system can generate designs that minimize D-error for almost any imaginable research problem.

But in what way is minimizing expected D-error optimal? Optimizing D-error as applied to pooled choice models is not the same as minimizing error in standard hierarchical Bayes analyses of choices. D-error arose originally from McFadden's (1974) multinomial logit choice model as a robust measure of expected error. However, most available tests estimate errors in parameters for an aggregate logit analysis. While researchers have created designs that minimize errors in the complex models, a feasible approach to optimize designs for the error around individual level partworths from hierarchical Bayes (HB) designs has yet to be developed. That is, for any design, it is possible to simulate expected errors, but determining the best of countless possible designs for HB is not possible even if we could run simulations for years.

The fact that the wrong model is being optimized does not appear to matter in practice. Simulations I have done indicate that as long as the heterogeneity of the respondents is not too great, and the designs that individuals see are not too unbalanced, then designs with low D-error will have low mean square error estimating individual utilities.

A much more serious problem arises because the logit model assumes that the errors around the utility for each task do not depend on the characteristics of the choice set. Errors are expected to be identically, equally distributed across tasks. However, people will generate less reliable choices when tasks are difficult than when they are simple. For that reason, apparently paradoxical results may be obtained when D-optimal designs are applied to people. Below are three examples where D-error can produce less effective designs with real people. These involve choice designs that have utility balance, partial profiles, and pair comparisons, all of which alter the error levels for individuals.

WHAT IS WRONG WITH UTILITY BALANCE?

Utility balance is one of the four conceptual criteria for statistically efficient designs, along with orthogonality, level balance and minimal overlap (Huber and Zwerina 1996). It says that a choice task provides more information if the expected utility difference among the alternatives is small. The intuition is simple; there is very little information from utility-unbalanced choice if one alternative dominates all the others in the set.

While utility balance means setting a prior partworth vector when developing a design, our experience is that there are few gains relative to setting all priors to zero. Several years ago Rich Johnson and I, along with a number of coauthors, experimented with adding utility balance in Sawtooth Software's CBC choice generation algorithm. We took the initial choice sets and then swapped one, two or three attribute levels to achieve more utility balance. With simulated respondents we achieved improvements in efficiency of 20%, implying that utility balance would deliver the same precision with 20% fewer respondents. We were quite excited.

However, when we ran those utility balanced designs against actual respondents there were no appreciable gains either in hit rates or predictions of choice shares for holdout stimuli. The reason is simple. Although there is more information in the utility balanced choices, these choices are, by definition, more difficult for respondents. The result of that difficulty was an increase in the error level and that reduced the precision of the estimates.

Thus, it is not that utility balance is wrong; it remains technically correct. It's just that it comes at a price. Unlike simulated respondents, humans react to utility balanced designs by making larger errors, and that error can negate any statistical advantage. I continue to recommend that conjoint designers use moderate levels of non-zero priors in developing their designs, but that serves largely to minimize the likelihood of dominated alternatives in a design. From a predictive perspective utility balance appears to make little difference.

WHAT IS RIGHT ABOUT PARTIAL PROFILES?

In a partial profile design, respondents see subsets of attributes that differ across choice tasks. Exhibit 1 gives an example from a paper by Keith Chrzan (2010). In the partial profile design respondents experience tasks in which each focuses on different subsets of the seven attributes. In contrast, in the full profile laptop condition the brands differ on all seven attributes. The full profile design displays minimal overlap in the sense that the alternatives have no overlap on any dimension.

We can expect behavioral differences between partial and full profile designs. Partial profiles require less effort simply because 6 pieces of information change in each task compared with 18 for the full profile. That simplicity will result in more accurate predictions if the simplification in partial profiles reflects what goes on in the marketplace. Consider first simplification. The time to respond to a full profile task begins at around 45 seconds but after 4-5 tasks quickly asymptotes to around 15 seconds. What we learn from the full profile task is the identity of the two to three attributes that drive these later and faster decisions. In contrast, the partial profiles compel attention to less important attributes. In the example partial profile shown in Exhibit 1, the speed of the CPU may not be an important in a full profile task but is more likely to emerge in the partial profile task because respondents are encouraged to think about it. Thus, we expect the variance in partworths across attributes to expand under full profile as more focus is given to more important attributes. By contrast, to the extent that lesser attributes are expected to be evoked in the market either by advertising, or promotional materials, then it will be important to see how they fare against more important attributes, and that will be better revealed by partial profile.

Exhibit 1
Full and Partial Profile Tasks

Source: Chrzan (2010)

Partial profile	***Option 1***	***Option 2***	***Option 3***
Operating System	**Windows 7 Home Premium**	**Apple/Mac OS X**	**Windows Vista Home Premium**
Processor (CPU)	3.2 GHz	1.6 GHz	2.6 GHz
Price	$1000	$800	$600

Full profile	***Option 1***	***Option 2***	***Option 3***
Operating System	**Windows 7 Home Premium**	**Apple/Mac OS X**	**Windows Vista Home Premium**
Processor (CPU)	3.2 GHz	1.6 GHz	2.6 GHz
RAM	2GB	4 GB	1 GB
Hard Drive	160 GB	250 GB	500 GB
Screen Size	13”	15”	17”
Battery Life	3.5 hours	5.5 hours	2 hours
Price	$600	$800	$1000

While there is some ambiguity about whether full profiles are more valid for markets than partial profiles, there is no doubt about the statistical cost of partial profiles. In the Chrzan case, the statistical efficiency of the partial design for a pooled logit design is half of the full profile. That implies that twice as many respondents would be needed to generate equal expected errors around the partworth values.

Chrzan’s surprising result is that the mean absolute error predicting full-profile holdout stimuli was 4% for partial and 6% for full profiles. Thus in this case, individual error overwhelmed the statistical efficiency of the design. This experiment is compelling because all models were built for pooled logit analysis, the appropriate theoretical context of D-error. However, Keith Chrzan indicated to me that the same predictive dominance by partial over full profiles occurs with an individual level analysis using hierarchical Bayes.

The purpose of this example is to demonstrate that there are conditions under which respondent error can overwhelm statistical error. However, it should not be seen as a general endorsement for partial profiles. Indeed, Jon Pinnell (http://www.sawtoothsoftware.com/download/techpap/pinnell.pdf) gives a number of examples (my own work included) where partial profiles produce greater prediction errors than full profiles. Thus, the gain from decreased error may or may not compensate for decreased efficiency, and where nonlinearities or interactions occur, then that loss of efficiency can be devastating.

Pair Comparisons for Difficult Choices

The following is an example of a task which is so difficult for respondents that only pairs are feasible. Typically, it is good to have more than two alternatives in a choice set. Indeed Johnson and Orme (1996, http://www.sawtoothsoftware.com/download/techpap/howmanyq.pdf) showed in a meta-analysis that 3-4 alternatives per task are optimal. Pairs are problematic because they are generally 25% less statistically efficient than triples and 35% less efficient than quads (see, Zwerina, Huber and Kuhfeld at http://support.sas.com/techsup/technote/mr2010e.pdf). Additionally, pairs can distort decision making because a respondent can simply count the number of ways one alternative beats the other and decide on that basis, effectively giving each attribute equal weight.

Exhibit 2
An Example of Binary Choice with
Two Attributes Differing

Choose a Procedure	Procedure A	Procedure B
Extent of the procedure	4 hour procedure 4 nights hospital stay 6 week recovery time	2 hour procedure 2 nights hospital stay 1 week recovery time
Risk of death during the procedure or within three years	2 deaths out of 100 patients	2 deaths out of 100 patients
Risk of heart attack during the procedure or within three years	1 heart attacks out of 100 patients	1 heart attacks out of 100 patients
Risk of stroke during the procedure or within three years	1 stroke out of 100 patients	1 stroke out of 100 patients
Risk of needing another procedure within three years	10 out of 100 need a new procedure	10 out of 100 need a new procedure
Change in life expectancy	Live 6 months longer	No change in life expectancy

The following study on people with heart conditions illustrates a case in which choice is so emotionally laden that the tradeoff must be in its most simple form. Respondents were asked to assume that to avoid an imminent heart attack they had to choose between surgical procedures that differed on six attributes: the extent of the surgical procedure, change in life expectancy, three-year risks of death, heart attack, stroke, and having to have another procedure. In our case, we pretested the task and found that triples or quads were not feasible for our respondents. Worse, initially they could not deal with more than two attributes differing.

We needed a design with 12 pairs having 6 attributes each at three levels that would have the following properties (see Exhibit 2). First, there would be six tasks for which only two attributes would differ. Then, there would be three tasks in which four attributes would differ, and finally, three tasks in which all six would differ. To avoid dominance or the majority of confirming dimensions heuristic, we required that each alternative would have as many attributes in its favor as against it. Thus, for tasks 7-9, there were four attributes differing, as in Exhibit 3. Generating such a design required Warren Kuhfeld's

SAS macros in which restrictions are put in the design tested. See http://support.sas.com/resources/papers/tnote/tnote_marketresearch.html .

Exhibit 3
Binary Choice with
Four Attributes Differing

Choose a procedure	Procedure A	procedure B
Extent of the procedure	3 hour procedure 3 nights hospital stay 4 week recovery time	2 hour procedure 2 nights hospital stay 1 week recovery time
Risk of death during the procedure or within three years	6 deaths out of 100 patients	2 deaths out of 100 patients
Risk of heart attack during the procedure or within three years	2 heart attacks out of 100 patients	2 heart attacks out of 100 patients
Risk of stroke during the procedure or within three years	1 stroke out of 100 patients	1 stroke out of 100 patients
Risk of needing another procedure within three years	10 out of 100 need a new procedure	20 out of 100 need a new procedure
Change in life expectancy	Live 1 year longer	Live 6 months longer

Exhibit 4 gives Warren's SAS code. The output from the code (not shown here) is valuable in a number of ways. First, it enables us to assess the efficiency cost of partial profiles. If respondents completed 12 pairs in which two attributes differed, the expected D-error would be .38. Our hybrid design decreases error by about a third to .26. However, 12 full profile pairs reduce it even more to .12. Second, it reveals weaknesses in the design through the expected variance-covariance matrix of the design parameters. In this case none of the variances of the parameters were out of balance. We ran with the hybrid design because we expected that the 2-attribute choices would prepare respondents for the more efficient 4 and 6 attribute choices.

Exhibit 4
SAS Code to Generate the 12 Tasks
Six with 2, Three with 4, and Three with 6 Attributes Differing

```
%mktruns(3 ** 6)

%mktex(3 ** 6, n=729)

%macro res;
  diff  = sum((x[1,] - x[2,]) ^= 0);      /* number of attributes differing*/
  favor = sum((x[1,] - x[2,])  > 0);     /*number that favor an alternative*/
  if setnum <= 6 then do;
    bad = bad + abs(diff - 2);
    if diff = 2 then bad = bad + abs(favor - 1);
    end;              /*Two attributes differ one favoring each alternative*/
  else if setnum <= 9 then do;
    bad = bad + abs(diff - 4);
    if diff = 4 then bad = bad + abs(favor - 2);
    end;                /*Four attributes differ two favoring each alternative*/
  else do;             /*Task numbers 10-12*/
    bad = bad + abs(diff - 6);
    if diff = 6 then bad = bad + abs(favor - 3);
    end;            /*Six attributes differ three favoring each alternative*/
  %mend;

%choiceff(data=design, model=class(x1-x6 / sta),
        restrictions=res, resvars=x1-x6, options=resrep relative,
        nsets=12, seed=104, flags=2, beta=zero)
```

We generated four designs using different random starts to the optimization routine. These had very similar efficiency scores, but differed strongly in terms of the actual choice tasks. The four design blocks serve two purposes. First, by pooling across these randomized blocks we can estimate almost any model (say with interactions) so we were not restricted to the main effects model. Second, it is easy to assess the reliability of the designs by comparing the pooled result across the blocks.

When applied to 225 respondents, we found that prediction to a full profile holdout was greater for the 6 choices differing on 2 attributes with 78% accuracy, compared to 67% accuracy when we restricted the hierarchical Bayes analysis to the last 6 choices. Thus the greater error in the multi-attribute tasks countered their clear statistical superiority. However, pooling all 12 tasks produced the highest hit rate with 81% accuracy. This result is consistent with lessened error from the simpler 2-attribute choices and a minor benefit from including the more complex ones. We also examined whether the different random blocks mattered. The average partworths from each block correlated around .80 with pooled result, indicating reasonable stability across blocks.

Four lessons can be drawn from this study. First, it again shows that greater accuracy can emerge from the less statistically efficient tasks in which only two attributes differ. Second, it demonstrates the way simpler tasks can prepare respondents to accept more complex ones. In this case respondents were first introduced to each attribute one at a time, then they made tradeoffs among pairs of attributes, and ultimately experienced more difficult tasks in which four or six attributes differed. Third, this example shows how statistical software can be used to generate relatively efficient designs with strong constraints. Finally, it demonstrates the way blocks can give the researcher a sense of the reliability of a given choice design.

One Choice Task per Respondent

Conjoint succeeds in revealing individual partworths precisely because it utilizes multiple responses per person. Therefore it seems strange to propose that one choice task per person would ever be appropriate. Such a context occurs for very difficult choice tasks that are not expected to be repeated or negotiated. These might include largely irreversible choices such as the decision to go to a particular college or one's first job. It also includes choices which have a large but uncertain impact on future welfare.

An example of one task per respondent comes from a national study of people's willingness to pay to limit arsenic in their tap water, where its presence can cause cancer and death for a handful of people out of 100,000. The task, shown in Exhibit 5, gives respondents the choice between a reduction in the coming year of the likelihood of death by 2/100,000 in return for an increase of $150 in their water bill. Notice this is framed as the kind of choice a water company might give to its customers. Although it was difficult, we did not find that people had difficulty answering the question. It was the follow up questions that were difficult.

If the respondent agreed to pay the $150 then the follow up increased the fee to $200. If they did not agree to the $150 then the follow up decreased it to $100. Instead of making a quick simple choice, the follow up choice requires a more difficult assessment of how much they are willing to pay.

We examined whether an analysis of the first choice would provide substantially different valuations than a group of iterative choices. We found that three iterations produced similar results as the initial choices, resulting in a value per statistical life of around $8 million. We also found that the relative weights and the influence of demographics on the choice were also similar, regardless of whether we used the initial choice or the final choice after the iterations. In this case, it appears that asking more than one choice per respondent made relatively little difference in the estimate of the statistical value of life.

Exhibit 5
Example of a Single Choice Study

> Assume for a moment that you have a **4 out of 100,000 chance** each year to eventually get bladder cancer and die due to arsenic in your drinking water.
>
> We are asking you to assume this is because arsenic exists at different levels in drinking water depending upon the region of the country and the type of water supply.
>
> What if your water treatment system could use a new treatment to reduce that risk to **2 out of 100,000 chance** each year?
>
> If you learned that this treatment increased your annual water bill by $150, what would you think of the new treatment?
>
> 1. I would be in favor of the new treatment and the increased water bill.
> 2. I would be opposed to the new treatment.
> 3. I have no preference for whether the new treatment is used or not.

The lesson here is that there are times when it is hard to do better than one choice per respondent. In that case, rather than vary the different choice tasks within each respondent, they can be varied across respondents. Of course, one loses the ability to estimate individual partworths. Additionally, one choice per respondent generally requires far more respondents to achieve stability for aggregate estimates. For example, in the arsenic study, we needed almost 1000 respondents to achieve strong statistical significance for our explanatory variables. By contrast, in other conjoint studies with 8-12 choices per respondent, aggregate partworths have been similarly stable with groups of 100 respondents.

WHEN SHOULD CHOICE DESIGNS BE HARD OR EASY?

This paper initially defined a horizontal divide between choice-based conjoint studies. The central idea is that respondents facing difficult choices need to have easier tasks. These easier choice designs have fewer alternatives per task, fewer attributes per alternative, and fewer tasks per respondent. The purpose of this section is to discuss three characteristics of the choice problem that force one to move to less statistically efficient but more respondent friendly designs. These choice characteristics shown in Exhibit 6 involve respondent’s emotions about the decision, their awareness of the attribute values, and their ability to make the required trade off.

Consider first the emotionality of the choices. Decisions are emotional when they carry substantial negative aspects that can instill regret or disappointment. Emotions also emerge for very important decisions, such as college choice or a home purchase, choices which involve large costs, long term implications, and substantial uncertainty. In such cases maybe there is a tendency to avoid the decision, to delegate the decision to the doctor, financial advisor or religious authority. In choice-based conjoint such decision aversion may be reflected in greater use of the ‘NONE’ or default alternative.

Exhibit 6
From Statistical Efficiency to Respondent Effectiveness
When to Move from Hard to Easy Choice Designs

CHARACTERISTICS OF THE CHOICE PROBLEM	**HARD CHOICE DESIGN:** **Many attributes, alternatives, choice tasks**	**EASY CHOICE DESIGN:** **Few attributes, alternatives, choice tasks**
EMOTIONALITY OF DECISION	Decision has clear objective criteria	Decision is difficult, emotional
AWARENESS OF ATTRIBUTE VALUES	Common attributes, benefits understood	New attributes, novel benefits
TRADEOFF FAMILIARITY, FREQUENCY	Similar decisions are repeated often	Decision is rare and typically one time

When emotion looms it is important to position choices within a familiar scenario. This has two benefits. First, it puts choices in a context where they might matter, rather than a purely hypothetical context that may seem like a game. Second, building a reasonable and credible story around the choice can increase the willingness to make the choice. We make choices all the time that we would prefer not to make, but one can learn from making difficult but potentially relevant choices. Finally, where emotion is involved it is wise not to include a default or “none” option. If a default is needed, it should be handled, as suggested by Karty and Yu (this volume), by a dual response where after the forced choice respondent can indicate the degree that the one chosen is acceptable in itself.

As in life, emotionality tends to be strongest in the initial choices, but then declines with repeated tasks. Thus, whether only a few or multiple tasks are appropriate depends on the purpose of the research. People adapt to difficult decisions both in reality and in conjoint exercises. Thus, if the goal is to see how people will respond after they have become accustomed to the decision, then multiple tasks are appropriate. By contrast, where the goal is to model the emotionality of one-time choices, then focusing the analysis on the first few tasks will be more likely to tap behavior in the thrall of that emotion.

The second characteristic of the choice problem is the extent that respondents know what the attribute means to them. Where attributes are familiar and people have experienced them, then decisions are relatively easy. By contrast, for unfamiliar attributes it is critical to invest in getting respondents to think about the attributes. Such familiarization of attributes occurs where there is a new technology or novel feature. In that case, it is valuable to encourage thinking about how such a novel benefit would be used. Sometimes the critical issue may be uncertainty about how others might respond to a product or service one might buy. In that case, a focus group discussing the attribute might be appropriate before the conjoint exercise. Finally, there are attributes where it is hard to assess how one feels about them. That problem was illustrated in the health example, where the different procedures could change ultimate life expectancy by six months. In our study we described that attribute and asked respondents to indicate on a 7-point scale how important that attribute is in their decision. We do not use that information in the analysis, but simply include the rating to help respondents prepare for a profoundly difficult but potentially important choice task.

The final characteristic of the choice problem is whether the respondent has familiarity making tradeoffs among the attributes. For example, one may have a clear sense of dread about 4 chances in 100,000 of dying from cancer from arsenic in tap water, and a sense of the value of $150 per year, but have no idea how to appropriately trade one off against the other. Probabilistic attributes, particularly for low probability outcomes, are especially difficult (for evidence see Johnson, Yang and Mohamed (this volume)). It is sometimes helpful to show a map with the number of deaths as red points, but typically people are insensitive to whether the denominator is 100,000 or 10,000. One way to avoid what is called ‘denominator neglect,’ is to frame the decision in terms probabilities that are between .2 and .8. Sometime this can be done by extending the time frame, to reflect lifetime chances, but in our case neither the frequency map nor aggregation was feasible. Instead, we gave respondents an idea of the magnitude of 100,000 people by comparing that quantity to a small city or a large stadium.

Where tradeoffs are not familiar, the survey can make them so. Begin with easy questions about the relative importance of each attribute compared with the others. Add some easy pairs where only two attributes differ to indicate which is more important. It is then possible to move to more complex choice tasks in which multiple attributes differ.

Summary and Conclusions

Three general conclusions derive from the previous analyses.

1. Make sure your design characteristics generate choices that respondents can handle. Choice design characteristics specify the number of alternatives per choice, the number of attributes differing, and the number of choice tasks per respondent. While these strongly impact statistical precision of the design, it is the ability of respondents to answer consistently that should determine these design characteristics.

It is critical to assess the way respondents react to the choices. For assessing familiar products, such as the price or positioning of well-known consumer goods, then the complexity of the choice can and arguably should match its complexity in the marketplace. Further, since the attributes are well known it does not make sense to prepare respondents for the choices, since the goal is to mimic relatively untutored behavior in the marketplace.

However, difficult choices with attributes requiring respondents to think deeply about their choices call for better preparation and simpler tasks. It may also limit the number of choice tasks each respondent can handle.

2. Once choice design characteristics are determined, use design optimization software to generate a good design. Minimizing D-error is very useful once the number of alternatives per task and the number of attributes differing has been established. By contrast, comparing D-error across different design characteristics can be deceptive. Examples given earlier showed where less statistically efficient partial profiles, pair choices and utility balanced designs provided greater predictive accuracy. However, given that the number of alternatives per choice and attributes per alternative have been established, then it is valuable to test the statistical efficiency of various designs. These can be done by the SAS system or through simulation in the Sawtooth Software system (via the *advanced design test*).

As mentioned earlier, it is not necessary to use priors to generate designs that take advantage of utility balance—assuming zero priors will work fine. The only reason to use non-zero priors would be to limit dominated alternatives, and that is only of concern in cases where there are only a few attributes differing in any choice set. Finally, randomization is important to ensure that unanticipated interactions do not distort the results. That randomization occurs automatically when Sawtooth Software generates many questionnaire versions and randomly assigns them to respondents. Where choice sets are prespecified, as with the SAS system, it is useful to generate a few independent versions to guard against mistakes and to assure that across respondents more general models are estimable.

3. Test, test, test. Do not believe me. It is certainly true that some of my assertions will not generalize across all product and service categories. It is important for all conjoint users to include variability in choice designs so that they can be tested. Johnson, Yang and Mohamed (this volume) demonstrated how to test the value of adding respondents. Further, it is possible to test whether few choice tasks would have worked better simply by rerunning the analysis with fewer tasks, and thereby testing the influence of the latter tasks on prediction.

With respect to the number of attributes differing in a choice task, consider varying the number of attributes that are the same and grayed out in a task. It is then possible to assess the extent to which statistically inefficient choices with high overlap predict better than full profiles.

The same tests can be done on the number of alternatives in each choice. Consider beginning with pairs and gradually moving to triples and quads. These progressive changes may make the choice task a little less boring and result in better predictions by introducing more difficult tasks later when respondents are better able to cope with them.

I close with a few words about the positive function of the horizontal and vertical divides in our field. Each divide defines different intellectual goals that speak to different audiences.   The vertical divide separates the elegance of optimal designs against the serviceability of simpler ones. For example, Sawtooth Software's balanced but randomized designs substitute robustness and general applicability, but bypass the intellectual puzzle that characterizes optimal design. Still, both sides benefit from the insights and procedures developed by the other, and that process is very well represented by the open debate and empirical focus among Sawtooth Software users.

Consider as well the horizontal divide that separates common but relatively trivial brand decisions against uncommon but critical life decisions. We naturally tend to believe that what works for our current studies applies to other, quite different studies. Just as we need to avoid claiming that there is a true utility value that is represented in our derived partworths, we need to be willing to accept that different kinds of problems require very different conjoint approaches. Here again, the breadth of ideas and methods presented in Sawtooth Software conferences continually reminds us of the multiple ways conjoint can be used to measure values, and should give us appropriate humility about the belief that the ways we do things now are best.

REFERENCES

Chrzan, Keith (2010) "Using Partial Profile Choice Experiments to Handle Large Numbers of Attributes," International Journal of Marketing Research, 52.6, 827-840.

Joel Huber and Klaus Zwerina (1996), "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research, 33* (August) 307-317.

McFadden, Daniel (1974), "Conditional Logit Analysis of Quantitative Choice Behavior," in Paul Sarenbka (ed), *Frontiers of Econometrics,* 105-142, New York: Academic Press.

ADAPTIVE BEST-WORST CONJOINT (ABC) ANALYSIS

[24]ELY DAHAN
[25]UCLA MEDICAL SCHOOL

EXECUTIVE SUMMARY:

Subjects choose the best and worst alternatives almost as easily as traditional choice-based conjoint where only the preferred option is chosen, as Finn and Louviere (1992) have shown, and with greater questioning efficiency (e.g. Orme 2005). Adaptive Best-worst Conjoint (ABC) utilizes full-profile questioning to instantaneously estimate utility for health care treatments at the individual level with an average of 13 tasks (11-17 range). A prostate cancer treatment application highlights the method and its benefits.

THE NEED:

While conjoint analysis has typically been used to help companies understand their markets, another potential application is self-help for individuals. For example, when facing alternative health care treatment choices, a patient might need guidance in prioritizing treatment aspects, health benefits, and side effects in making a final choice.

The medical self-help context raises additional requirements for conjoint implementation. For example, it would be helpful to obtain immediate preference estimates in-clinic and report them quickly in preparation for discussions between healthcare professionals and their patients. As each survey task is completed, utility estimates should be updated and the next question chosen adaptively to reduce interview time. And the user interface needs to be sufficiently easy for patients to use so that they don't give up. It needs to be readily available to and affordable within the healthcare setting. We employ a macro-based Excel worksheet to address the issues of instant estimation and reporting.

BACKGROUND:

To maximize efficiency and speed, we build on prior work in best-worst, or MaxDiff, questioning developed by Finn and Louviere (1992). Flynn et al. (2007) have highlighted efficiency and estimation benefits of having patients choose the best and worst out of a set of alternatives rather than the "pick one" task used in choice-based conjoint analysis (CBC). We use this as a starting point and add adaptive questioning and transitivity to speed things up even more.

Orme (2005) showed in a Sawtooth Software white paper that best-worst tasks with more than about 4 or 5 alternatives at a time didn't seem to offer much statistical benefit,

[24] Researcher, UCLA Medical School, 2542 Cardigan Court, Los Angeles, CA 90077-1377, elydahan@gmail.com

[25] This report describes research done within UCLA's Urology Research Group within the Department of Urology at UCLA's Geffen School of Medicine. This research was sponsored by a grant from the National Institute of Health under contract No. R01CA134997-01A1. Comments on the research from Dr. Chris Saigal, Dr. Robert Kaplan, Dr. Mark Litwin and Dr. Kate Crespi have been extremely helpful as has the participation of 41 subjects, and the research assistance of Sylvia Lambrechts and Elizabeth Garcia.

using synthetic MaxDiff data. Chrzan and Patterson (2006) showed actual respondents tasks with 3 items at-a-time, all the way through 8 items at-a-time. They observed only small differences in performance of the scores derived, and concluded:

> *"Given Orme's (2005) findings about increasing accuracy with larger numbers of MaxDiff questions, and our finding that task length increases with number of items per question, we recommend a larger number of questions with smaller numbers of items per question. Given the slight evidence of poorer hit rates and poorer out-of-sample for 3 items per question we recommend using 4 or 5 items per question in MaxDiff experiments."*

Based on both Orme (2005) and Chrzan and Patterson (2006), ABC tasks feature a 4 stimuli at-a-time format. The comparison of ABC to this prior research is not entirely apples-to-apples as those authors tested standard MaxDiff, where each item was a single statement (i.e. one attribute), whereas ABC utilizes full-profile conjoint stimuli that include bundles of attributes.

Of course, nothing inherent in the ABC algorithm limits its use to a specific number of stimuli at a time, so it remains an empirical question what the "optimal" number of stimuli to be shown in each task should be. In fact, ABC could even allow different numbers of stimuli, starting with three, for example, and building up to five, as the survey proceeds and the respondent learns the task.

ABC's best-worst algorithm is based on reconstructing a rank order for full-profile conjoint analysis, building on the work of V. Seenu Srinivasan, Paul Green and many others. The ranking of the *n* full-profile stimuli derives directly from individually identifying the *n* x (*n*-1) / 2 paired comparisons through adaptive best-worst questioning and transitivity.

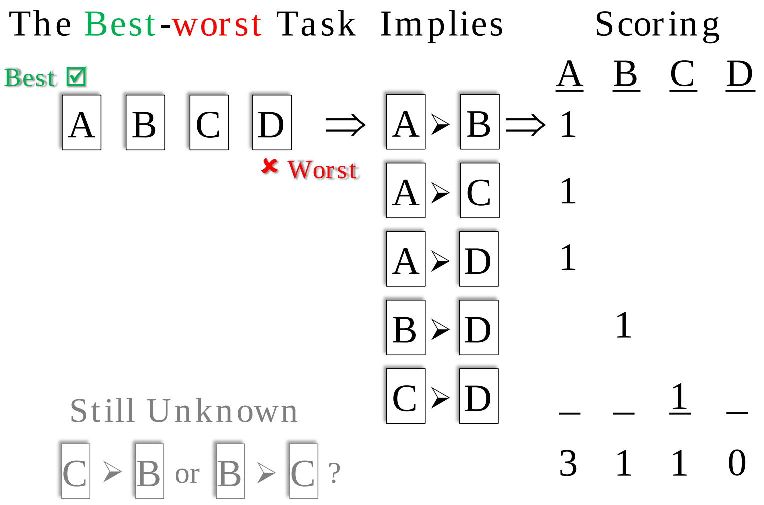
Prior attempts at adaptive conjoint questioning include Sawtooth Software's Adaptive Conjoint Analysis (ACA), Sawtooth Software's Adaptive Choice-Based Conjoint (ACBC), Toubia, et al's FastPACE, and Dahan's Conjoint Adaptive Ranking Database System (CARDS). It is worth noting that Fraenkel, et al. (2001 and 2004) successfully utilized Sawtooth Software's Adaptive Conjoint Analysis (ACA) to survey patients and generate utility reports.

We build on prior research by speeding up data collection and report generation using a new adaptive approach and transitivity of preference. Unlike prior approaches, ABC adaptively poses new best-worst tasks based on resolving the greatest number of yet-to-be tested paired comparisons in each task. ABC also utilizes transitivity to resolve paired comparisons when preferences may be inferred from already-resolved paired comparisons without the need for a direct response.

With the ABC approach, utility functions with 7-10 parameters can be estimated instantaneously at the individual level with as few as 12-15 tasks completed in 10 minutes.

While CBC with four options at-a-time produces only three paired comparisons (**1st** Choice > option **B**, **1st** Choice > option **C**, **1st** Choice > option **D**), ABC Best-worst with four options generates *five* of the six possible paired comparisons (**Best** > option **B**, **Best** > option **C**, **Best** > **Worst**, option **B** > **Worst**, option **C** > **Worst**; only **B** is not compared to **C**). So ABC is 66% more efficient than CBC in resolving paired comparisons, even without adaptive questioning.

Figure 1

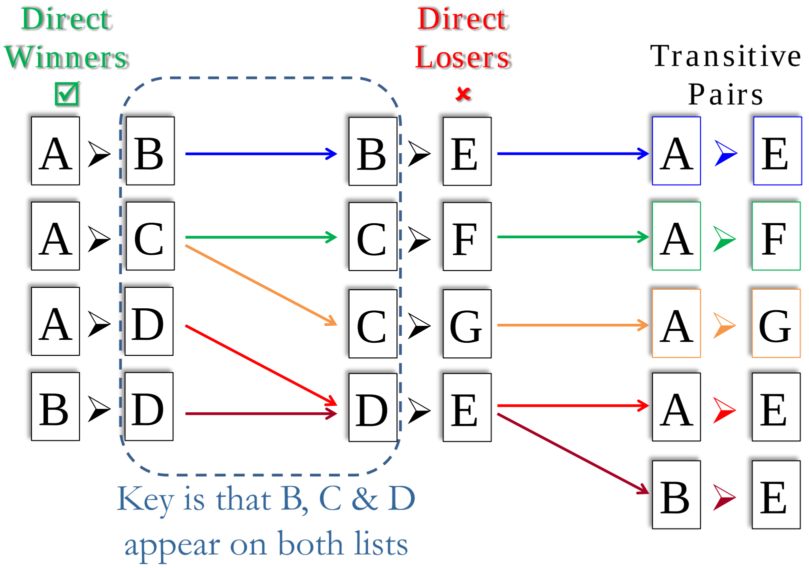


The figure above illustrates how choosing full-profile “A” as best and “D” as worst resolves five of six possible paired comparisons. It also shows how “winning” a paired comparison adds to that full-profile’s “score.”

This inherent efficiency advantage of best-worst questioning is further enhanced through adaptive questioning based on transitivity of preference. That is, we assume that if full-profile **A** is preferred to full-profile **B**, and if **B** > **E**, then **A** is also > **E**, even though we never directly compared **A** to **E**.

Figure 2

Transitivity: How are transitive "wins" revealed?



In the above figure, eight paired comparisons that were previously resolved through direct questioning also resolve five *new* paired comparisons that had not yet appeared in direct questioning. Such transitivity may resolve even more paired comparisons than direct questioning. For example, with 16-full-profiles, there are 16 x 15 / 2 = 120 possible paired comparisons. ABC typically requires 12 best-worst questions to estimate a utility function with 7-9 parameters. Each of the 12 tasks directly resolves 5 paired comparisons (though some redundancy is likely, so they may not all be unique pairs), so direct questioning resolves at most 60 pairs (5 x 12). The remaining 60+ paired comparisons are resolved through transitivity.

Transitivity's contribution grows with higher number of full-profile stimuli, making ABC quite scalable. Once all possible paired-comparisons have been resolved directly or through transitivity, each full-profile can be scored based on how many comparisons it won, as shown in the figure below for our 16 full-profiles. If a full-profile both wins and loses its paired comparison, it is scored as a tie and earns only 0.5 wins.

Figure 3

Methodology: Paired Comparisons
Count up all of the paired comparison results

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1		0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1		1	0	0	0	0	1	1	0.5	1	0	0	0	0
4	1	0	0		0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1		1	0	0	1	1	1	1	1	0	0	1
6	1	1	1	1	0		1	0	1	1	1	1	1	0	0	1
7	1	1	1	1	1	0		0	1	1	1	1	1	0	0	1
8	1	1	1	1	1	1	1		1	1	1	1	1	0	0	1
9	1	1	0	1	0	0	0	0		1	0	0	0	0	0	0
10	1	1	0	1	0	0	0	0	0		0	0	0	0	0	0
11	1	1	0.5	1	0	0	0	0	1	1		1	0	0	0	0.5
12	1	1	0	1	0	0	0	0	1	1	0		0	0	0	0
13	1	1	1	1	0	0	0	0	1	1	1	1		0	0	1
14	1	1	1	1	1	1	1	1	1	1	1	1	1		1	1
15	1	1	1	1	1	1	1	1	1	1	1	1	1	0		1
16	1	1	1	1	0	0	0	0	1	1	0.5	1	0	0	0	
Sum	**15**	**13**	**8.5**	**14**	**4**	**4**	**4**	**2**	**11**	**12**	**8**	**10**	**6**	**0**	**1**	**7.5**

Doctor and Family Support this treatment	Sex: Same as before treatment
Active: Treatment requires action within weeks	Urinary: No problems
No Cutting: Treatment does NOT require any surgery	Bowel: No problems
	Lifespan: Live my expected lifespan

1

Doctor and Family Support this treatment	Sex: Decreased compared to before treatment
Active: Treatment requires action within weeks	Urinary: Long-term issues
Cutting: Surgery with some risks and hospital time	Bowel: Short term urgent & frequent bowel movements
	Lifespan: Live 5 years fewer than expected

14

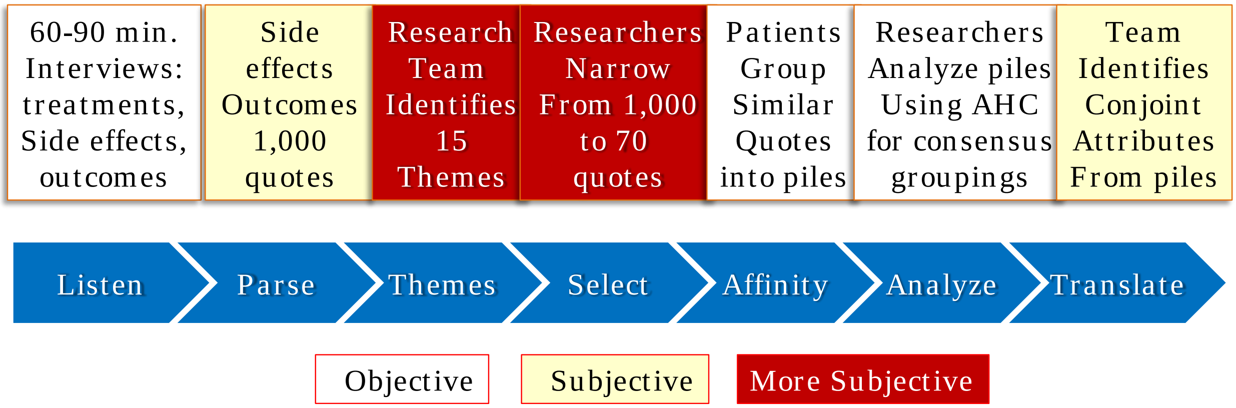
The Adaptive element of ABC is based on choosing the four stimuli to show in each best-worst task based on maximizing the number of unresolved pairs. Whether this strategy is actually better than strategies based on increasing the precision of utility estimates is not known. The question merits further research.

Methodology Overview

The first step in implementing ABC is to determine the attributes and levels that encompass the individual's key considerations in making a choice. We identified these for our empirical test using Dahan and Saigal's (2012) "Voice of the Patient" (VoP, based on Griffin and Hauser's 1993 "Voice of the Customer" research) as shown in the figure below.

Figure 4

“Voice of the Patient” (VoP) Process


60-90 min. Interviews: treatments, Side effects, outcomes
Side effects Outcomes 1,000 quotes
Research Team Identifies 15 Themes
Researchers Narrow From 1,000 to 70 quotes
Patients Group Similar Quotes into piles
Researchers Analyze piles Using AHC for consensus groupings
Team Identifies Conjoint Attributes From piles
Listen
Parse
Themes
Select
Affinity
Analyze
Translate
Objective
Subjective
More Subjective

Sample quotations from the VoP process are shown in Figure 5.

Figure 5

VOP Attributes: Sample narratives from men treated for prostate cancer

Treatment Issues

Cutting: I don't want to be cut; I don't want to have surgery.

Others' Advice: I only follow doctors’ advice up to a point. Not 100%

Caution: I could wait for a while if the numbers stay stable…

Action: I was just thinking "we have got to do something"

Side Effects

Sex: If you have an understanding partner, the ED thing can be ok.

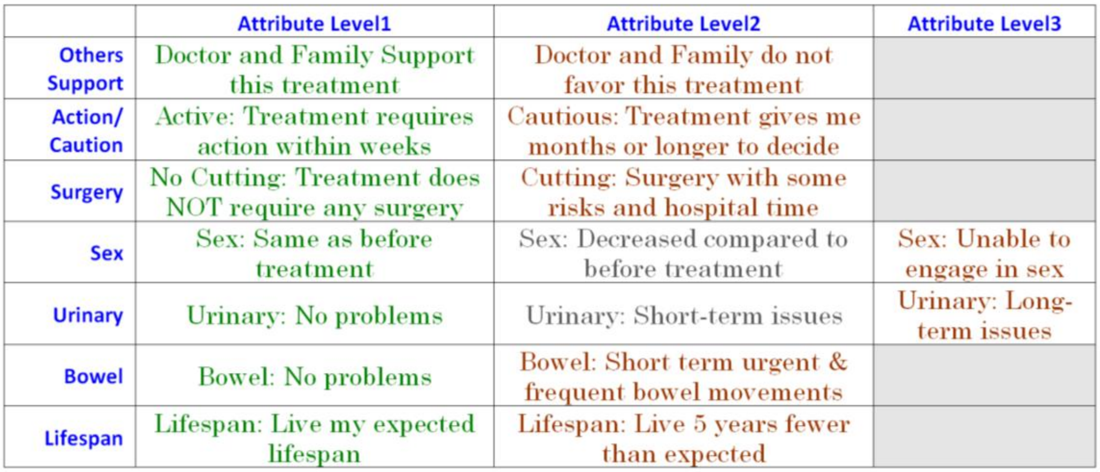
Urinary: Changing pads frequently…feels as if you don't have control of your life.

Lifespan: It is more important to stay alive, regardless of the side effects.

Bowel: The bowel issue is the biggest deal because it is socially unacceptable.

Based on the Voice of the Patient process, we identified the seven attributes and associated levels shown in Figure 6.

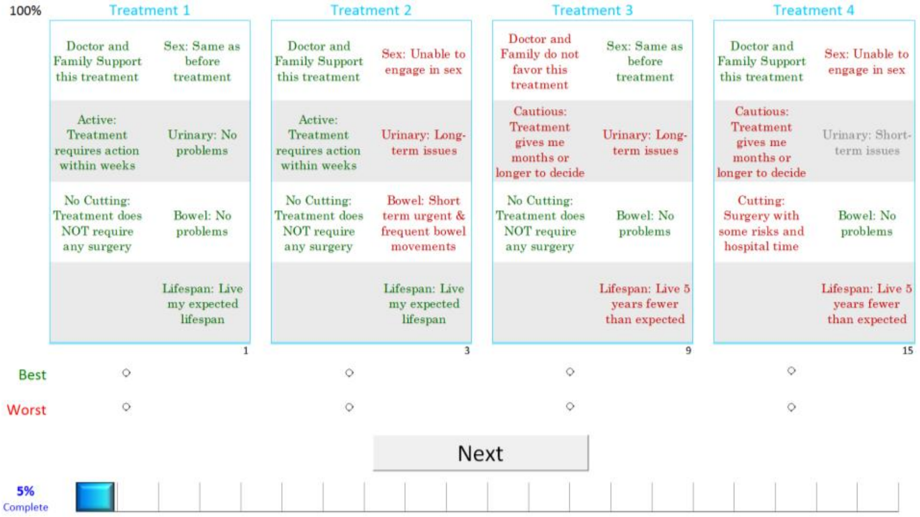
Figure 6: Seven Attributes with 2-3 Levels Each based on the Voice of the Patient



	Attribute Level1	Attribute Level2	Attribute Level3
Others Support	Doctor and Family Support this treatment	Doctor and Family do not favor this treatment	
Action/ Caution	Active: Treatment requires action within weeks	Cautious: Treatment gives me months or longer to decide	
Surgery	No Cutting: Treatment does NOT require any surgery	Cutting: Surgery with some risks and hospital time	
Sex	Sex: Same as before treatment	Sex: Decreased compared to before treatment	Sex: Unable to engage in sex
Urinary	Urinary: No problems	Urinary: Short-term issues	Urinary: Long-term issues
Bowel	Bowel: No problems	Bowel: Short term urgent & frequent bowel movements	
Lifespan	Lifespan: Live my expected lifespan	Lifespan: Live 5 years fewer than expected	

The seven attributes in Figure 6 are then converted into an orthogonal, balanced design of 16 full-profile conjoint stimuli to be displayed as part of the adaptive best-worst conjoint survey. A sample best-worst task is depicted below in Figure 7.

Figure 7: Adaptive Best-worst Conjoint (ABC) question format



100%

Treatment 1	Treatment 2	Treatment 3	Treatment 4
Doctor and Family Support this treatment Sex: Same as before treatment	Doctor and Family Support this treatment Sex: Unable to engage in sex	Doctor and Family do not favor this treatment Sex: Same as before treatment	Doctor and Family Support this treatment Sex: Unable to engage in sex
Active: Treatment requires action within weeks Urinary: No problems	Active: Treatment requires action within weeks Urinary: Long-term issues	Cautious: Treatment gives me months or longer to decide Urinary: Long-term issues	Cautious: Treatment gives me months or longer to decide Urinary: Short-term issues
No Cutting: Treatment does NOT require any surgery Bowel: No problems	No Cutting: Treatment does NOT require any surgery Bowel: Short term urgent & frequent bowel movements	No Cutting: Treatment does NOT require any surgery Bowel: No problems	Cutting: Surgery with some risks and hospital time Bowel: No problems
Lifespan: Live my expected lifespan	Lifespan: Live my expected lifespan	Lifespan: Live 5 years fewer than expected	Lifespan: Live 5 years fewer than expected
1	3	9	15

Best ○ ○ ○ ○

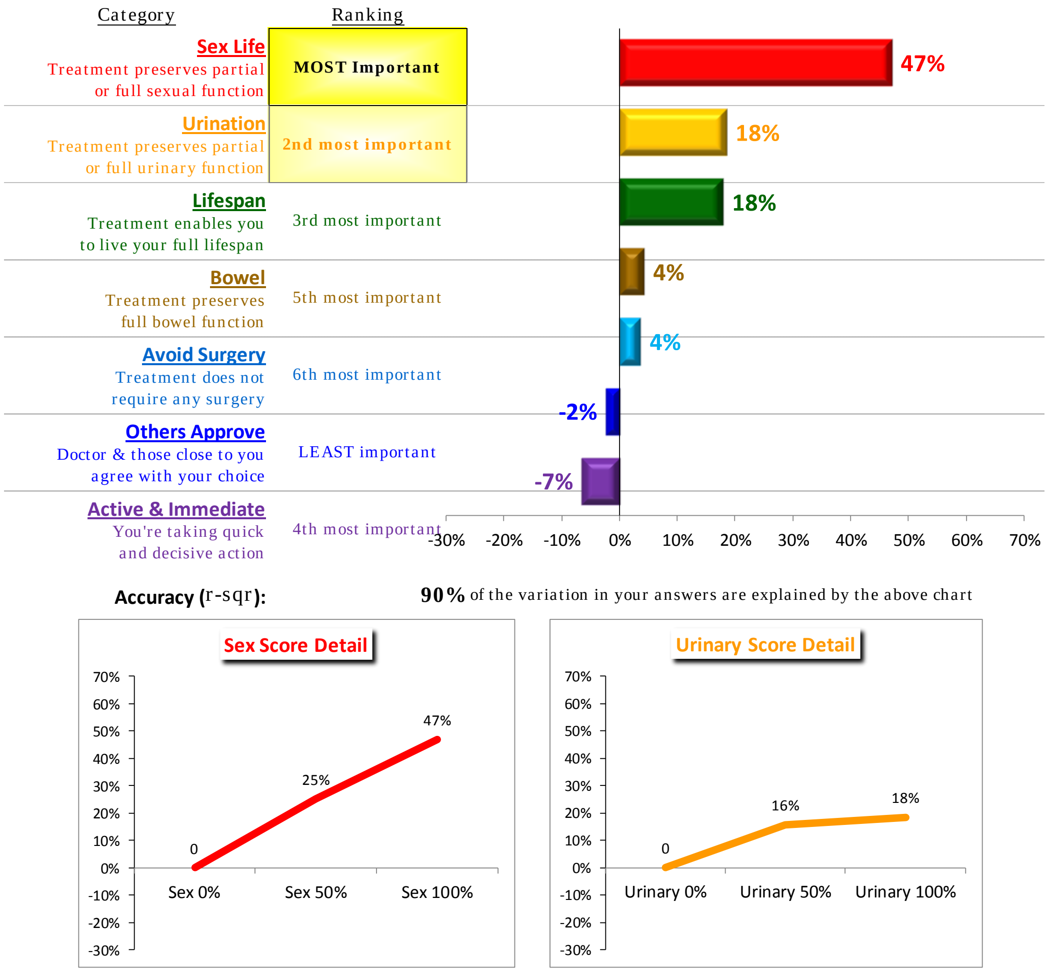
Worst ○ ○ ○ ○

The entire application has been built as an Excel spreadsheet file with macros. Respondents are first taught about attributes and levels, then primed with a sample task, then guided through 12 or more such choice tasks among four full-profiles at a time. The full profiles can be text- or graphics-based. In the background, the Excel software tracks every possible paired comparison and records all of the direct comparisons as well as calculating all of the transitive ones.

The paired comparisons are converted into scores for each full-profile card, and a utility function estimated based on these scores.

A key goal of ABC in healthcare is to improve the conversation between patients and health professionals. An individualized report, a sample of which is shown in Figure 8, is automatically generated at the end of the ABC process so that the individual can learn about his or her personal preferences. The report serves as the basis of discussion for post-biopsy meeting between patient and medical professional.

Figure 8: ABC Output: A Personalized Report for Treatment Priorities



Category	Ranking
Sex Life Treatment preserves partial or full sexual function	MOST Important
Urination Treatment preserves partial or full urinary function	2nd most important
Lifespan Treatment enables you to live your full lifespan	3rd most important
Bowel Treatment preserves full bowel function	5th most important
Avoid Surgery Treatment does not require any surgery	6th most important
Others Approve Doctor & those close to you agree with your choice	LEAST important
Active & Immediate You're taking quick and decisive action	4th most important

Accuracy (r-sqr): **90%** of the variation in your answers are explained by the above chart

Empirical Test: Real Patients Making Treatment Decisions

Led by principal investigator Dr. Chris Saigal under NIH Research Contract R01CA134997-01A1, ABC was tested in clinic against two traditional methods of patient preference assessment. Dr. Saigal has graciously allowed us to share some initial findings from this research.

We conducted a randomized trial with prostate cancer patients from the Los Angeles Veterans Hospital and UCLA Medical facilities comparing the effectiveness of ABC conjoint analysis versus the more traditional time trade off and rating scale methods (see Figure 9 and Figure 10).

We recruited men at the VA urology clinic and UCLA Urology Clinic undergoing prostate needle biopsy for suspicion of prostate cancer.

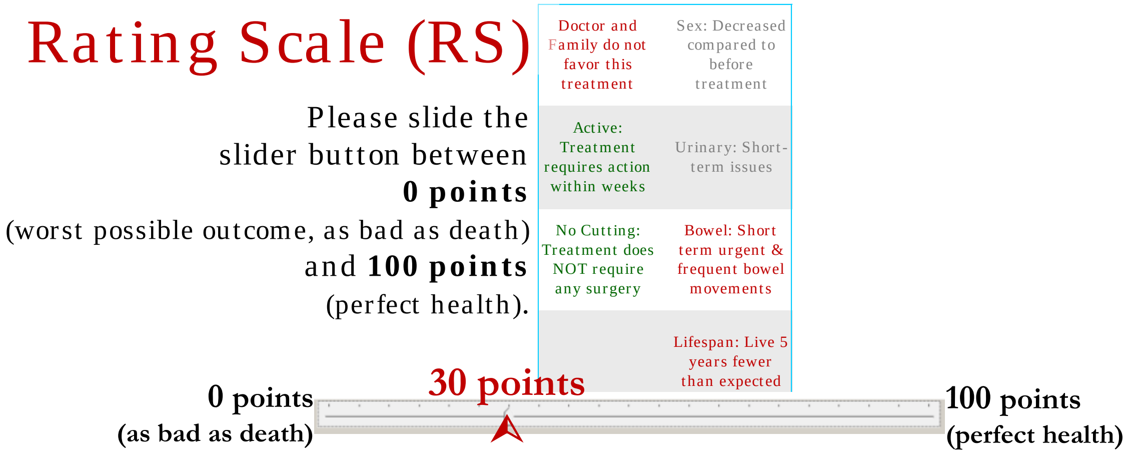
The Rating Scale (RS) and Time Trade Off (TTO) applications were developed utilizing the same stimuli as the conjoint study, but implemented with software pre-designed to collect data for those two traditional methods.

The Adaptive Best-worst Conjoint (ABC) approach, in contrast, was developed from scratch to address the need for rapid estimation and reporting at the individual patient level.

Subjects and task order randomized to one of two conditions:

- Rating Scale (RS)

Figure 9: Rating Scale (RS) Task



Rating Scale (RS)

Please slide the slider button between **0 points** (worst possible outcome, as bad as death) and **100 points** (perfect health).

Doctor and Family do not favor this treatment	Sex: Decreased compared to before treatment
Active: Treatment requires action within weeks	Urinary: Short-term issues
No Cutting: Treatment does NOT require any surgery	Bowel: Short term urgent & frequent bowel movements
	Lifespan: Live 5 years fewer than expected

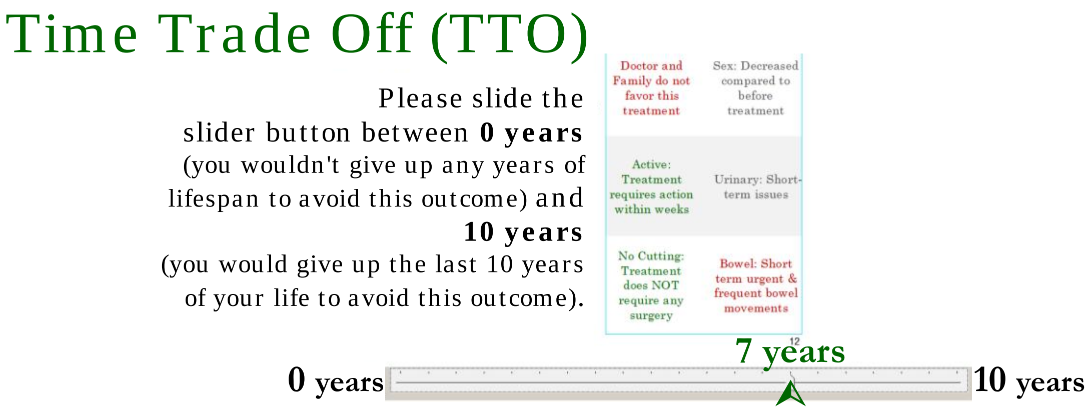
30 points

0 points (as bad as death)

100 points (perfect health)

- Time Trade Off (TTO)

Figure 10: Time Trade Off (TTO) task



Time Trade Off (TTO)

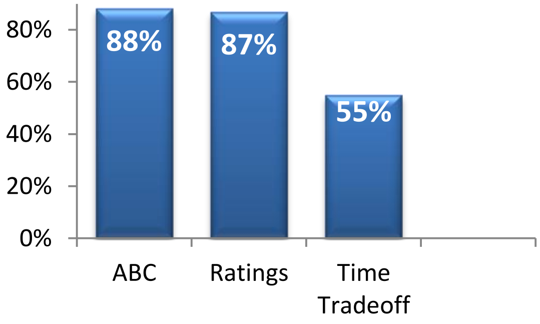
Please slide the slider button between **0 years** (you wouldn't give up any years of lifespan to avoid this outcome) and **10 years** (you would give up the last 10 years of your life to avoid this outcome).

Doctor and Family do not favor this treatment	Sex: Decreased compared to before treatment
Active: Treatment requires action within weeks	Urinary: Short-term issues
No Cutting: Treatment does NOT require any surgery	Bowel: Short term urgent & frequent bowel movements

KEY RESULTS:

- **Internal consistency**: ABC had the highest degree of internal validity as measured by the r-squared between estimated and actual utility scores for each of the 16 stimuli. Rating Scale was not far behind, but TTO seemed to produce less internally consistent responses. The mean r-squared's for the three methods are shown in Figure 11.

Figure 11: Mean R-Squared between estimated and actual scores for the 16 stimuli



- **Holdout prediction**: In addition to the RS/ABC or TTO/ABC tasks, subjects also completed two holdout tasks that required them to rank order four stimuli. The first holdout utilized 4 randomly chosen holdouts from the original 16, while the second holdout task utilized never-before-seen stimuli, but in the same format, from a different experimental design. Figure 12 shows the first-choice hit rates averaged across subjects and holdout tasks. In addition to ABC, RS and TTO, a hierarchical Bayesian analysis of the data, performed at Sawtooth Software, was conducted and used to predict holdout responses. The results show that ABC with simple multiple linear regression outperforms the other two methods, and does even better with HB estimation.

Figure 12: First choice hit rate (25% randomly guessed correctly)
Consistent with the above result, Figure 13 analyzes the same holdout questions, but now considers all possible paired comparisons rather than just first choice.

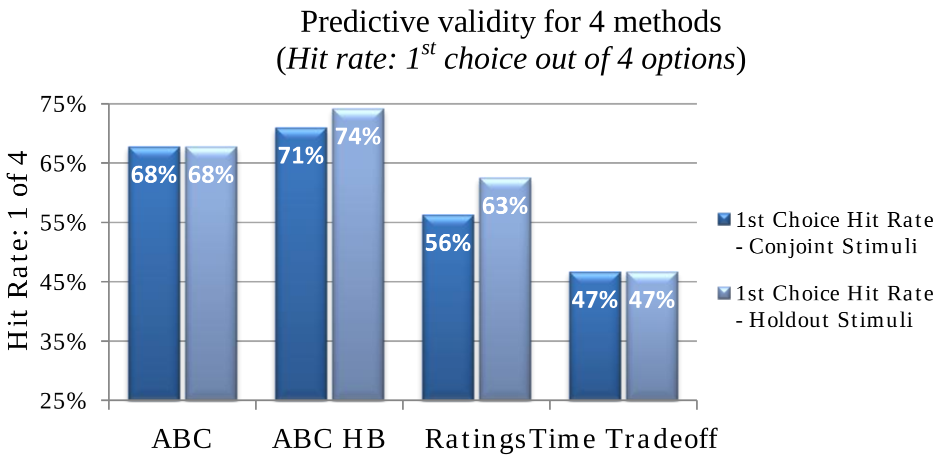
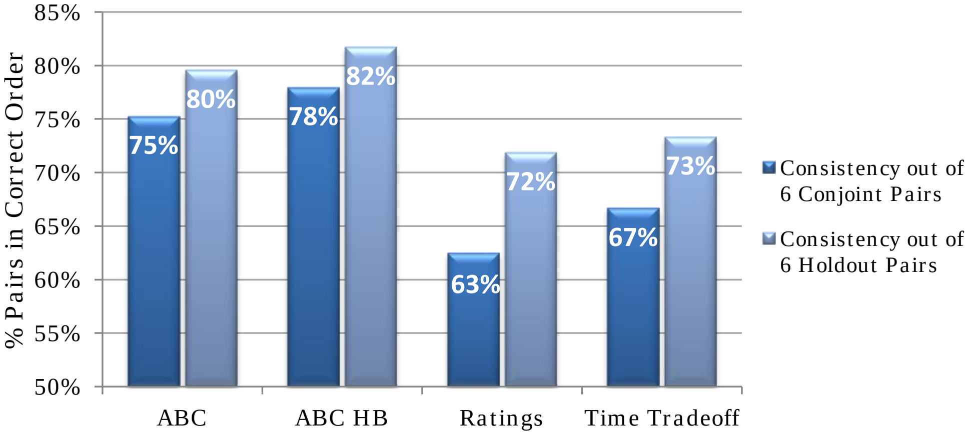


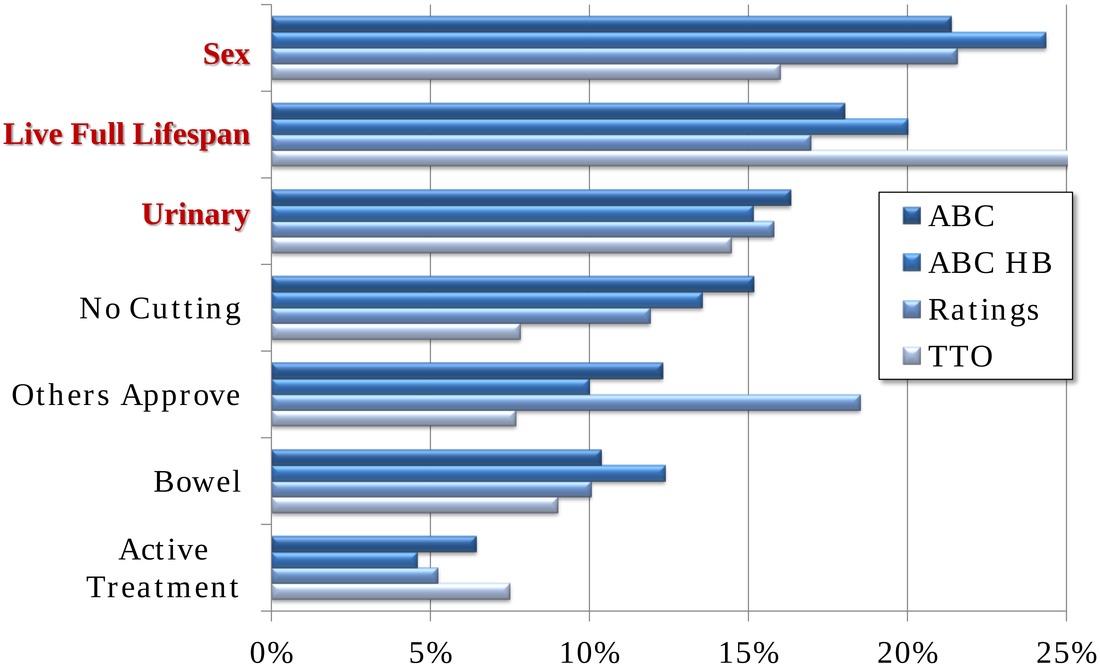
Figure 13: Pairwise Consistency scores for four methods
(*% of pairs in the correct order*)



Once again, ABC outperforms the other two methods, and HB improves prediction.

Figure 14 demonstrates a high degree of consistency across all methods regarding the most and least important attributes. The exception is that TTO seems to over-emphasize the importance of a 5-year change in life span relative to the other methods, possibly since it emphasizes life span as the unit of measure of utility for all tasks.

Figure 14: Most important attributes



- **Respondent satisfaction**: Based on survey results thus far, both patients and doctors agree that communication between them has been facilitated by discussing the conjoint analysis report generated based on ABC. Further, patients who utilized ABC are reporting a higher level of satisfaction with the treatments they chose than those who did not utilize ABC.

METHOD LIMITATIONS:

Scaling Up: One limitation of ABC is the need to limit the number of stimuli, and hence the number of parameters that can be estimated.

As we add more parameters beyond the (9) nine estimated in the empirical test, we need ~ 2 more stimuli per extra parameter. For example, if (16) cards is sufficient to estimate (9) parameters, we might need (27) - (32) cards for (18) parameters. But comparing four full-profile stimuli in each ABC task, each with 12-15 attributes depicted, could easily overwhelm respondents. So there is an upper limit as to the number of parameters that can be estimated at the individual level due to cognitive limitations. This is true of virtually all conjoint methods. A solution may be to start with many *possible* attributes and levels, but to narrow or prioritize them, individual-by-individual, prior to starting the ABC task.

Also, while (16) cards have 16 x 15 /2 = 120 pairs to resolve, (32) cards have 32 x 31 / 2 = 496 pairs to resolve. The number of paired comparisons to be resolved grows as the square of the number of stimuli.

The potentially solution may lie in the fact that transitivity also resolves more pairs as the number of cards grows.

If this is not enough, we can bring in other approaches working in the background, such as Dahan's (2007) CARDS or other logical methods of resolving pairs automatically.

Two other limitations of ABC, as presently designed, are:

- The lack of intensity of preference measurements, i.e. A is assumed to be preferred to B by the same amount as B is preferred to C in each paired comparison
- The issue of estimating rank ordered data using multiple linear regression as opposed to more sophisticated statistical or operations research methods such as multinomial Logit (with or without hierarchical Bayesian post hoc analysis), LINMAP, or maximum likelihood estimation.

SUMMARY:

Theoretical

- ABC is efficient due to best/worst questioning *and* transitivity
- Estimation methods are many, but even lowly multiple linear regression in Excel seems to perform reasonable well
- Scaling up to more attributes and levels may be feasible by combining ABC with adaptive methods such as CARDS or other automated pair resolution methods
- There may be ways to incorporate Individual ABC with HB-based "Crowd Wisdom" to achieve better decisions
- As suggested by Rich Johnson at the Conference, the algorithm for adaptively selecting the next four tasks to be shown could incorporate not only unresolved paired comparisons, as is done now, but also uncertainty in the parameter estimates

Managerial

- Individual conjoint is markedly different from market conjoint
- In order to diffuse ABC, a "Plug-N-Play" Excel Template may be the place to start
- More empirical research needs to be done as the present data represents a small sample
- ABC builds on prior work by Louviere, Flynn, Fraenkel and others, and represents a novel method of measuring patient preferences in real time, with immediate reporting using adaptive best-worst conjoint analysis.

References

Chrzan, Keith and Michael Patterson (2006), "Testing for the Optimal Number of Attributes in MaxDiff Questions," Software Research Paper, www.sawtoothsoftware.com/download/techpap/mdoptimalatts.pdf, 7 pp.

Dahan, Ely and Christopher Saigal (2012), "Voice of the Patient," 2012 Sawtooth Software Conference presentation.

Finn, A., Jordan J. Louviere (1992), "Determining the appropriate response to evidence of public concern: the case of food safety," *Journal of Public Policy and Marketing*, 11, pp. 12–25.

Flynn, Terry N., Jordan J. Louviere, Tim J. Peters and Joanna Coast (2007), "Best–worst scaling: What it can do for health care research and how to do it," *Journal of Health Economics*, 26:1, pp. 171–189

Fraenkel, Liana, Sidney Bogardus, and Dick R. Wittink (2001), "Understanding Patient Preferences for the Treatment of Lupus Nephritis with Adaptive Conjoint Analysis" *Medical Care*, 39:11, November, pp. 1203-1216.

Fraenkel, Liana, S. T. Bogardus, J. Concato, D. T. Felson, D. R. Wittink (2004), "Patient preferences for treatment of rheumatoid arthritis," *Annals of the Rheumatic Diseases*, 63:pp. 1372–1378.

Orme, Bryan (2005) "Accuracy of HB Estimation in MaxDiff Experiments," Sawtooth Software Research Paper, www.sawtoothsoftware.com/download/techpap/maxdacc.pdf, 7 pp.

Maximizing Purchase Conversion by Minimizing Choice Deferral: Exploring the Impact of Choice Set Design on Preference for the No-Choice Alternative

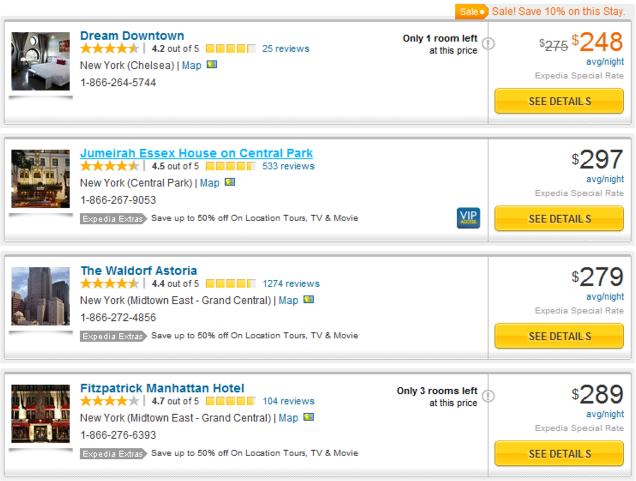
Jeffrey P. Dotson
Vanderbilt University
Jeff Larson
Brigham Young University
Mark Ratchford
Vanderbilt University

Introduction

In this paper we adopt the perspective of a marketplace or retailer and consider the impact of product assortment (i.e., real world choice sets) on preference for the no-choice alternative. Whereas brand-centric studies are focused on determining the impact of own- and competitive- marketing activities on sales for a particular brand, retailers are less concerned with brand performance and are more concerned with the overall performance of the category. In other words, they care less about which specific product is selected and more about whether a consumer makes a purchase.

For example, online marketplaces like Expedia, Orbitz, and Travelocity are primarily interested in maximizing purchase conversion or, alternatively expressed, are interested in minimizing choice deferral. Figure 1 illustrates the type of choice tasks produced by online websites. They are given limited space to describe an often times large collection of products using limited information. Given the competitive nature of the industry they must present the alternatives in a compelling enough form to encourage consumers to make a purchase. Decision variables include the collection of attributes presented, the form of presentation, and the order in which products are presented.

Figure 1. Illustration of choice task for an online hotel marketplace



Sale Sale! Save 10% on this Stay.

Dream Downtown
4.2 out of 5 25 reviews
New York (Chelsea) | Map
1-866-264-5744
Only 1 room left at this price
$~~275~~ $248 avg/night
Expedia Special Rate
SEE DETAILS

Jumeirah Essex House on Central Park
4.5 out of 5 533 reviews
New York (Central Park) | Map
1-866-267-9053
Expedia Extras Save up to 50% off On Location Tours, TV & Movie
VIP ACCESS
$297 avg/night
Expedia Special Rate
SEE DETAILS

The Waldorf Astoria
4.4 out of 5 1274 reviews
New York (Midtown East - Grand Central) | Map
1-866-272-4856
Expedia Extras Save up to 50% off On Location Tours, TV & Movie
$279 avg/night
Expedia Special Rate
SEE DETAILS

Fitzpatrick Manhattan Hotel
4.7 out of 5 104 reviews
New York (Midtown East - Grand Central) | Map
1-866-276-6393
Expedia Extras Save up to 50% off On Location Tours, TV & Movie
Only 3 rooms left at this price
$289 avg/night
Expedia Special Rate
SEE DETAILS

In preparation for this research we analyzed data from an online travel website to determine the extent to which consumers defer choice. We examined data for one of the largest DMAs in the US for a two-week period and found that close to 15,000 unique searches were conducted for hotels in that market. Of those 15,000 searches, only 123 resulted in the actual purchase of a hotel room. This corresponds to a conversion rate of less than 1% and is, to our understanding, typical of conversion rates in the industry. Examination of this data raises the obvious question: why do so many consumers elect to defer choice? Although there are many possible explanations (e.g., they are not in the market, are learning about the alternative, are in the process of preference formation, etc.) our attention in this paper is focused one very specific cause that is formally referred to as choice deferral.

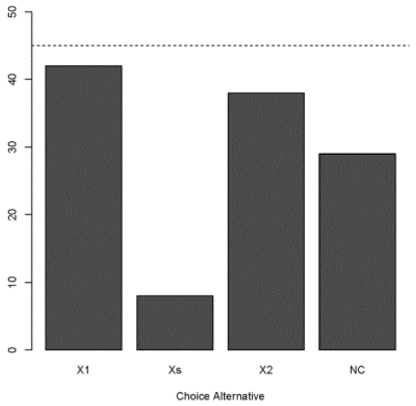
Choice deferral is a construct that has been studied extensively in the marketing literature (Dhar, 1997). In its essence, the theory of choice deferral suggests that in unforced choice settings (i.e., choice tasks that include the no-choice alternative), consumers will be more likely to defer a decision if the choice is cognitively difficult. Specifically, consumers will be more likely to defer choice if the alternatives in the choice set are more, rather than less, similar to each other. Our goal in this paper is to develop and estimate a theoretically based model that explains and can capture the effects of choice deferral. By so doing, we can learn how to optimize the presentation of products in an assortment to minimize choice deferral and maximize purchase conversion.

Model Development

Our model is grounded both theoretically and conceptually using models of evidence accumulation. Specifically, we base our model formulation on the results of the Dependent Poisson Race Model (DPRM) (Ruan et al., 2007) and a multivariate normal approximation to the same (Dotson et al., 2012). According to the DPRM and other “race models” of choice, consumer choice results from the stochastic accumulation of evidence over time. When faced with a choice task, consumers amass evidence in favor of alternatives in a choice set as their attention attenuates between the various alternatives and their corresponding attribute levels. This evidence arrives according to known stochastic process and is recorded on alternative specific "counters". The first alternative in the choice set to accumulate some minimally sufficient level of evidence (i.e., the alternative the crosses the finish line first) is the one selected by the decision maker.

Figure 2 provides a simple illustration of this process for a choice set with two alternatives (denoted X1 and X2) and a no-choice option (denoted NC). For a moment, ignore the bar labeled Xs. The bars plotted for X1, X2, and NC represent, respectively, the evidence counters for those alternatives. The dashed line denotes the finish line for the race. At each discrete unit of time *t* a signal is received and is recorded for each of the alternatives. If the alternative is attractive, it will receive (on average) large signals relative to less attractive alternatives. For example, in the illustrated race, alternative X1 leads both X2 and NC in terms of its accumulated evidence so far. This process of evidence accumulation continues until an alternative crosses the finish line or threshold, meaning it is chosen by the respondent.

Figure 2
An illustration of the a race model with a no-choice alternative



Using this process as the guiding theoretical framework, we begin by specifying a race model that accumulates evidence according to a Poisson process with rate λ_i, where *i* denotes the alternative in the choice set. Ultimately we will connect the conjoint design to the choice process by expressing $\lambda_i = f(x_i)$, where x_i denotes the vector of attribute levels that comprise the design of alternative *i*. The no-choice alterative is incorporated into the model by adding it as an additional option with a known and fixed rate, λ_0. Given the properties of a Poisson distribution, we can write the expected utility and variance of alternative *i* as:

$$E(U_i) = Var(U_i) = T\lambda_i$$

where *T* denotes the length of the race. Note that although the most attractive alternative (i.e., the one with the largest λ_i) will always be chosen in expectation, this is not true for a particular realization of a race, as the evidence accumulation process is intrinsically stochastic.

Following the DPRM, similarity between alternatives is modeled by decomposing the rates for each alternative into a component that is unique to the alternative and a component that is shared between a pair of alternatives, $\lambda_i = \lambda_{ui} + \lambda_{ij}$. In this expression λ_{ui} is the component that is unique to alternative i and λ_{ij} is the component that is shared between alternatives i and j. This dependence allows us to write the covariance between alternatives i and j as:

$$Cov\left(U_i, U_j\right) = T\lambda_{ij}$$

The theoretical effect of this dependence is also illustrated in Figure 2. Although we omitted it from discussion earlier, the bar for Xs denotes an accumulator for the evidence that is shared between alternatives 1 and 2. Shared evidence arises when alternatives share the same level of a given attribute (e.g., two hotels have the same star-level). As shared evidence does not distinguish either of the two alternatives under consideration, its effectively slows the race down. That is, it takes longer for the alternatives with common attribute levels to reach the finish line. This is the key property of this model as it relates to choice deferral. Because the no-choice alternative accumulates evidence at a fixed rate that is independent of the remaining alternatives, its choice probability will increase if the remaining items are similar to each other. In other words, if the no-choice alternative is still running the race at the same rate while the other alternatives are slowing down it is more likely that it will win.

Although conceptually appealing, the DPRM presents a variety of challenges to estimation, particularly with complex choice designs that have many attributes and/or alternatives. We follow Dotson et al. (2012) and develop a more simplistic approach to estimation by exploiting the asymptotic relationship that exists between a dependent Poisson process and a multivariate normal distribution. This is similar to the relationship that exists between the univariate Poisson distribution and a univariate normal distribution. Specifically, as the collection of λ_i become large we can write the distribution of utility for all alternatives in a given choice set as:

$$U \sim N(\lambda, \Sigma)$$

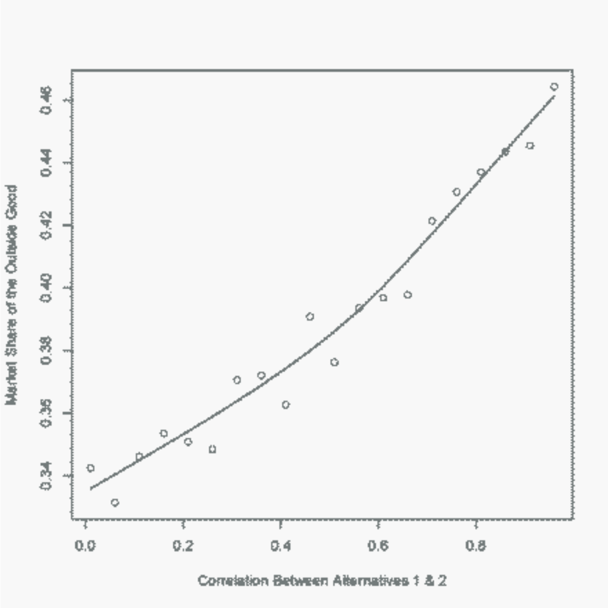
where U denotes the vector of utilities for all alternative in an arbitrary choice set of size J+1 (the +1 allows us to incorporate the no-choice alternative), λ is the mean vector of utility that is parameterized as a function of the design of alternatives and Σ is a covariance matrix that captures dependence among the alternatives in the choice set, and is of the form:

$$\Sigma = \begin{bmatrix} \lambda_1 & \cdots & \lambda_{1,J} & 0 \\ \vdots & \ddots & \vdots & 0 \\ \lambda_{J,1} & \cdots & \lambda_J & 0 \\ 0 & 0 & 0 & \lambda_0 \end{bmatrix}$$

Note that the model above results from the fact that the mean and variance of the Poisson distribution are equal to each other. Taken collectively, this specification leads us to a multinomial Probit model (MNP) with a structured covariance matrix, where the structure is derived from the properties of the DPRM. Estimation can then proceed using standard Bayesian techniques developed for the MNP. We allow for respondent heterogeneity by fitting an HB MNP model. Discussion of the estimation algorithm is omitted here, but we refer the interested reader to Dotson et al. (2012) for details.

To illustrate the benefits of our proposed model, consider the simulation results presented in Figure 3. In this simulation, we examine a choice set of 2 alternatives and a no-choice option. We compute the probability of picking the no-choice alternative (i.e., the expected share of the outside good) as a function of the similarity (or cross-alternative correlation) of alternatives 1 and 2, where similarity is computed using our proposed model. As the similarity of the alternatives in the choice set increases, our model predicts that the probability of picking the outside good will also increase. This is exactly what is implied by the literature on choice deferral, thus providing validation that our model is capturing behavior that is relevant to the conversion problem faced by marketplaces and retailers.

Figure 3
Relationship between product similarity and preference for the no-choice alternative



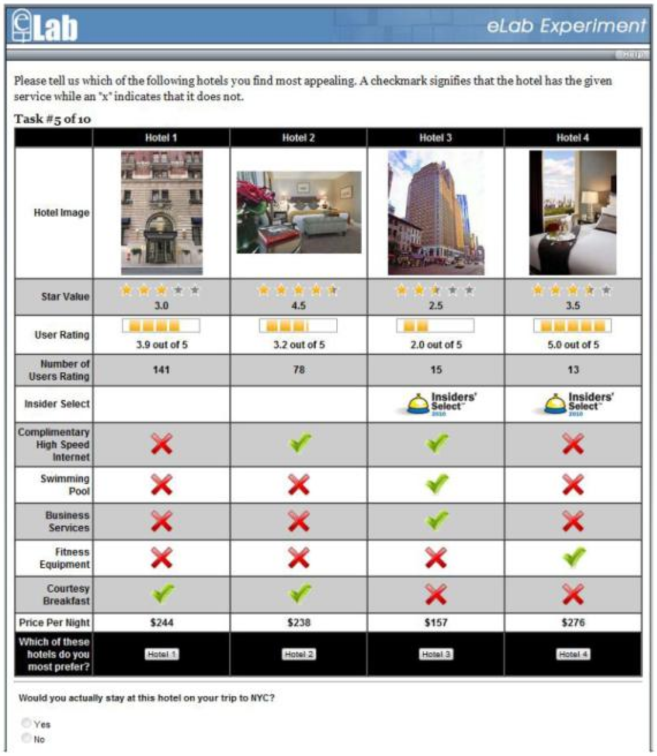
It is important to note that the effect illustrated in Figure 3 is not a prediction that emerges from standard choice models (i.e., the heterogeneous multinomial logit model - MNL), where the relationship between the similarity of alternatives in the choice set and the preference for the outside good are assumed to be independent. Moreover, the MNL predicts that as the choice set expands in size (i.e., as we move from 4 to 5 alternatives) the share of the outside good will decrease, irrespective of the design of the additional product(s). Our model, however, suggests that the addition of an additional product(s) into the choice set will be detrimental unless the new product(s) is perceptually distinctive from those in the existing choice set.

Discrete Choice Study

To test our proposed model we designed a conjoint study to assess preference for hotels in the Midtown Manhattan market in New York City. This market was selected because it is a geographically compact area with a large number of hotels that offer substantial variety in terms of their objective and perceptual similarity. The study was fielded using respondents from Vanderbilt University's eLab Consumer Panel. The sample was restricted to include individuals who intended to take a trip to New York City in the coming year, yielding a total of 219 usable responses.

Subjects completed 20 choice tasks like the one illustrated in Figure 4. We utilized a dual-response design to allow respondents to express preference for the no-choice alternative, while simultaneously maximizing the information contained in each completed choice task. In a dual response choice task respondents first indicate which of the alternatives presented they find most attractive and are then asked if they would actually purchase their preferred option. If not, they are treated as having chosen the no-choice alternative. This approach provides us with a partial rank of alternatives (i.e., increases the information content of the choice task) and helps avoid some respondents' tendency to artificially avoid giving no-choice answers so as not to seem uncooperative. A full list of attributes included in the study appears in Table 2. In addition to verbally described attributes, we also included hotel image as an attribute in the study. All images were pretested using in a separate study where respondents provided a rating of the appeal of the image, thus allowing us to include the average appeal as a covariate in the utility function.

Figure 4
Example choice task from our discrete choice study



Please tell us which of the following hotels you find most appealing. A checkmark signifies that the hotel has the given service while an "x" indicates that it does not.

Task #5 of 10

	Hotel 1	Hotel 2	Hotel 3	Hotel 4
Hotel Image				
Star Value	★★★☆☆ 3.0	★★★★★ 4.5	★★★☆☆ 2.5	★★★★☆ 3.5
User Rating	3.9 out of 5	3.2 out of 5	2.0 out of 5	5.0 out of 5
Number of Users Rating	141	78	15	13
Insider Select			Insiders' Select 2010	Insiders' Select 2010
Complimentary High Speed Internet	✗	✓	✓	✗
Swimming Pool	✗	✗	✓	✗
Business Services	✗	✗	✓	✗
Fitness Equipment	✗	✗	✗	✓
Courtesy Breakfast	✓	✓	✗	✗
Price Per Night	$244	$238	$157	$276
Which of these hotels do you most prefer?	Hotel 1	Hotel 2	Hotel 3	Hotel 4

Would you actually stay at this hotel on your trip to NYC?

○ Yes
○ No

Results

To provide a basis of comparison, we contrast the results of our proposed model with a model commonly used in practice, the Hierarchical Bayesian Multinomial Probit (HB MNP) model with an Identity covariance matrix. Although parameter estimates will differ in scale, the HB MNP will yield parameter estimates that are inferentially identical to the model most commonly used in practice, the HB Multinomial Logit model. Table 1 presents in-sample and out-of-sample fit statistics for both models. In-sample fit is characterized using the Log Marginal Density, where a larger (i.e., less negative) number is indicative of better fit. Our proposed model yields a superior LMD of −5,843.7 compared to the HB MNP model LMD of −6,253.1. We contrast fit in a hold-out sample of 1 observation per respondent using both the hit rate (i.e., the average choice probability for the hold-out choice) and hit count (# of individuals for whom we are able to correctly predict their choice). Our proposed model produces a superior hit rate (0.48) and hit count (111) than the HB MNP (0.44 and 108).

Taken collectively, we can conclude that our proposed model is superior in terms of both explaining variation in the choice data and predicting behavior out-of-sample. This improvement in fit is wholly attributable to the non-zero cross-alternative covariance matrix included in the model.

Table 1
Model Fit Statistics

	LMD	Hit Rate	Hit Count
Model	**(In-sample fit)**	**(Out of Sample Fit)**	
HB Multinomial Probit Model w/ Identiy Covariance Matrix	−6,253.1	0.44	108
Proposed Model	−5,843.7	0.48	111

The posterior mean of the draws of $\bar{\beta}$ (i.e., average coefficients estimates taken across respondents) appear in Table 2. Although the estimates differ in scale, they are directionally consistent. The one notable exception is the estimate of the coefficient for the attribute, “Star Level,” which is positive for the HB MNP model and negative for our proposed model.

Table 2
Coefficient Estimates for the Discrete Choice Model

Attribute	HB MNP	Proposed Model
Price	−1.40	−0.71
User Rating	0.37	0.52
# of Ratings	0.01	0.12
Star Level	0.06	−0.08
Insider Select (Best in Class Award)	0.11	0.18
Internet	0.80	0.95
Pool	0.46	0.52
Business Services	0.16	0.21
Fitness	0.36	0.47
Pets Allowed	−0.08	−0.07
Breakfast	0.52	0.69
Visual Appeal	0.14	0.14

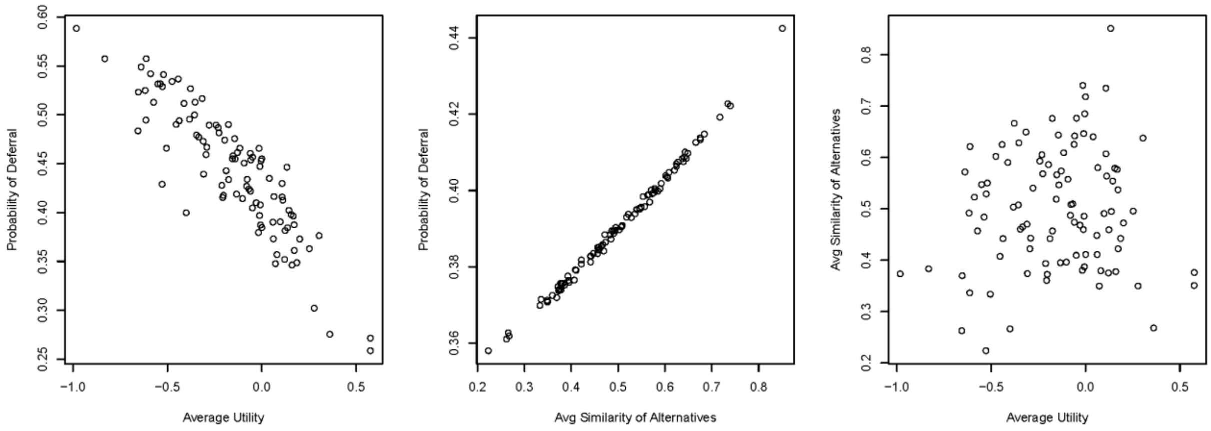
Although the fit statics presented in Table 1 provide compelling evidence in favor of our proposed model, the real benefit of our approach is best illustrated using a simulation study. Results of this study are presented in Figure 5. To produce these results we drew random subsets of hotels from the population of 83 hotels in the midtown Manhattan geographical region. For all hotels in given subset we computed the implied similarity using the approach described above. We also compute the average utility of each subset of hotels and the corresponding probability of choice deferral (i.e., the probability of picking the outside good). This simulation exercise was repeated many times and the results are presented in Figure 5.

The left panel of Figure 5 plots the average utility of each subset of hotels against the probability of choice deferral. As expected, the resulting relationship is negative. As the average attractiveness of a set of hotels increases, respondents are less likely to pick the no-choice alternative and are more likely to make a purchase. This implies that one strategy that can be employed by a retailer or marketplace to increase purchase conversion is to present consumers with better, more attractive options.

The center panel in Figure 5 plots the relationship between the average similarity of choice alternatives in each random subset of hotels and the probability of choice deferral. As predicted by the theoretical work on choice deferral, there is a strong positive relationship between the similarity of choice alternatives and the probability of choice deferral. As the hotels in the choice set become increasingly similar to each other, it is more difficult for respondents to identify the best option, thus motivating them to opt out of the choice. This result suggests that retailers can increase purchase conversion by presenting consumers with alternatives that are well differentiated.

Finally, the right panel in Figure 5 plots the average utility of alternatives in each random subset against the average similarity of alternatives. As illustrated, there is no discernible relationship between the two variables. This is an important finding as it demonstrates that these two strategies to decreasing choice deferral (i.e., increasing average attractiveness or decreasing similarity) can be applied independently of one another, thus providing retailers with two viable approaches that can be used to influence consumer choice.

Figure 5
Simulated relationship between product attractiveness, similarity, and probability of choice deferral



Conclusion and Implications for Practice

In summary, we find evidence that if alternatives in a choice set are similar to each other in design, consumers will be more likely to select the no-choice alternative, even if an otherwise utility maximizing alternative is available. We show that this behavior can be explained, both theoretically and empirically, using a model of evidence accumulation. We develop and estimate a model based upon the dependent Poisson race model and show that our model can be used to better forecast preference for the outside good, thus allowing us to optimize the presentation of alternatives in the choice set. Taken collectively, the findings from our study suggest two key implications for discrete choice practitioners:

First: Choice context matters! Although this point has been extensively demonstrated by researchers in behavioral economics and marketing, the models routinely used in discrete choice applications do not account for the many of the influences of context on choice. We highly recommend that practitioners use estimation and simulation techniques that can accommodate these effects. At a minimum, model estimation should account for heterogeneity in consumer preference through the use of hierarchical Bayesian methods. Simulation based techniques like Randomized First Choice are also useful in helping capture more realistic forms of consumer behavior. When required, a variety of specialized techniques like the one presented in this paper have been developed that can deal with specific effects of context on choice.

Second: Designing real world choice tasks is very different from designing a good conjoint choice task. Whereas the latter are most informative when respondents are confronted with challenging choices that force them to trade-off valued attributes, presenting consumers with difficult choices in the real world will lead to choice deferral. Although this statement may seem self-evident, consumers in online markets are routinely given tools that allow them to refine the results of their search, thus increasing the similarity of presented alternatives (e.g., sorting hotels by price or star level). The result of this refinement is likely to be an increase choice deferral. The results of our model suggest that it may be optimal to provide consumers with a reasonable mix of differentiated products, thus allowing them to easily discern which of the products best suits their needs. In other words, we should make it as easy as possible for consumers to make a choice.

A limitation of our existing study is that our model is constructed using conjoint data collected from choice sets of size 4. In practice, consumers are exposed to choice sets that contain potentially many alternatives. Although our proposed model can be scaled to include many alternatives, it is not clear that this approach will accurately depict the way consumers will choose. Further research is required to better understand how the size of the choice set influences our proposed choice process.

REFERENCES

Dhar, R. (1997). "Context and Task Effects on Choice Deferral." Marketing Letters 8(1): 119-130.

Dotson, Jeffrey P., Peter Lenk, Jeff Brazell, Thomas Otter, Steven MacEachern, and Greg Allenby (2012), “A Probit Model with Structured Covariance for Similarity Effects and Source of Volume Calculations,” Working Paper (contact first author for a copy).

Ruan, Shiling, Steven MacEachern, Thomas Otter and Angela Dean (2008), “Dependent Poisson Race Models and Modeling Dependence in Conjoint Choice Experiments”, Psychometrika, 73, 2, 261-288.

MENU-BASED CHOICE MODELING (MBC): AN EMPIRICAL COMPARISON OF TWO ANALYSIS APPROACHES

PAOLO CORDELLA[26]*,*
CARLO BORGHI,
KEES VAN DER WAGT,
AND GERARD LOOSSCHILDER
SKIM GROUP

ABSTRACT

We empirically compare a variation of the "sampling of alternatives" model to analyzing menu-based choice modeling data with the "serial cross-effects" model—one of the models that may be applied by Sawtooth Software's MBC. The main reason for looking into alternative models was to reduce the "art" component in the mixture of art and science that is needed to conduct MBC studies. With the Sampling of Alternatives method the cross-effects are not formally included, as it is a main-effects model. However, because individual-level models are built via HB, and due to the fact that feasible (likely) alternatives are all being considered as alternatives within a competitive set, then the substitution between alternatives as offered by the logit model can help account for substitutability among items on the menu that are viewed as similarly preferred. With serial-cross effects modeling, one has to identify and include all relevant cross-effects him/herself. As with all processes, the more manuals steps needed, the more prone to errors it becomes.

However, if it is unlikely that many cross effects occur, or if they are easy to detect, we suggest using the serial cross effects model of Sawtooth Software's MBC. That's because it is readily available, while the sampling of alternatives model requires extensive custom preparation.

INTRODUCTION

In the early nineties, the concept of customization was introduced to describe how, in product categories like computers, automotive products, apparel or restaurant food, brands allow customers to custom-build their product. The idea is that customization drives satisfaction. In automotive, buyers can personalize their car by choosing options and accessories at an additional cost. In computers, Dell has made it part of its business model to custom order a personalized configuration, so that the result exactly meets the need. In fast food, it is common to allow customers to assemble their own menu from a variety of options on display over the counter. In fact, customization has become so successful that we now witness a trend in the opposite direction: simplifying consumer choices by offering bundled options. The customer can choose from a predefined subset of bundled options, trading off freedom of choice against a discount or the certainty of the match. For example, in apparel, dress dolls are dressed up to suggest matching items. Bundles are common in telecom, where it is an effect of converging markets. In telecom we are moving from triple play (phone, cable and internet) to quadruple play (phone,

[26] Corresponding author: P.O. Box 29044 | 3001 GA Rotterdam | The Netherlands**, T** +31 10 282 3568 | **E** p.cordella@skimgroup.com.

cable, internet and mobile). The new idea is that the task simplification drives satisfaction, and customers are incentivized for the reduction of product line complexity.

The market research world has also embraced the idea of customization by introducing build your own (BYO) exercises, for example in menu-based choice modeling (MBC). The value of build-your-own tasks in MBC is complementary to the "traditional" full product profile approach as employed by Choice-Based Conjoint (CBC). CBC replicates consumer choice behavior in an environment of fully specified products, whereas MBC replicates consumer choice behavior in an environment of custom-built products (like Dell's).

Menu-based Choice Modeling exercises deliver item-level forecasts of market performance, by delivering:

- A feature or item "willingness to pay" assessment, delivering demand curves on an item level among many items: Forecast revenue and find the optimal price for all items on the menu, measuring uptake and supporting a decision whether or not to add a new item to a portfolio of items. The models help to assess cross-effect price sensitivity and cannibalization effects: does decreasing the price of single items hurt full menu sales? They also help us identify the most often chosen combinations and their prices, suggesting which items to bundle and insight into budget constraints, and how many items can we stuff in a bundle before we exceed the decision maker's budget?
- Insight into the performance of Mixed Bundling, e.g., what is the portion of customers still buying the individual items (the "and" choice) when they are also offered a matching bundle (including some items) at a discount (the "or" choice). It represents a disjunctive choice situation of choosing either a pre-configured bundle or *a la carte* options.
- Similarly, we can model conjunctive choices, were we offer customers a choice of base products or platforms and the opportunity to add zero to n "attachments" or options. At SKIM we call this "optional choice modeling," a variation to menu-based choice modeling. For example, it is used to determine if customers would be willing to purchase additional insurance with their initial product (e.g., travel insurance with a trip, or accessories with a car). In optional choice modeling we may even go as far as that the opportunity of buying the options affects the choice of the initial product (e.g., I only buy a telecom package if it offers me the right choice of handsets).

As practitioners, we have a history of looking into menu-based choice modeling approaches. We have developed our own models, based on repeated BYOs and optional choice models. We presented a variation of the "sampling of alternatives" model at the 2011 SKIM/Sawtooth Software Event. At the time, our main interest was validating the methodology by looking at measures such as predictive validity. The launch of the Sawtooth Software menu-based choice module offers the opportunity of an empirical comparison. This paper empirically compares this variation of a "sampling of alternatives" model with the "serial cross-effects" model that is possible (among other models) to specify within Sawtooth Software's MBC program, in a study into consumer choices of the features of notebook computers.

A REVIEW OF ANALYSIS APPROACHES

Academic attempts to model BYO type of choices using Menu-Based Choice Models (MBC) date back to the prominent work of Ben Akiva and Gershenfeld (1998). They presented a nested logit model structure to model multi-feature choices in which the buy/no-buy option is the upper level and a choice among possible alternatives to purchase is the lower level. Other academic research investigated alternative model structures, including multivariate probit (Liechty *et al.*, 2001). Most literature is based on Random Utility Maximization theory, which assumes that an agent faces a finite set of alternatives and chooses the one that maximizes her utility function. Key variations are based on which choice set to consider, and what choice structure to take to model the choices that the agent makes.

Two main theoretical models are the Exhaustive Alternatives models (EA) and the Serial Cross Effects model (SCE). Reviewing both models pointed to drawbacks that we thought to solve by using a variation of the Exhaustive Alternatives Model, which we call the Choice Set Sampling model (CSS).

SERIAL CROSS-EFFECT MODEL (SCE)

In the Serial Cross Effects model (SCE), the choice of each item is modeled separately. The dependent variable is the single choice of the item. As a result, we need to estimate as many different logit models as we have items in the study. Each logit model is used to predict the choice of item X as a function of its own price (own effect), and potentially other items' prices (cross-effects). In the presence of cross effects, it means that the choice of item X is dependent on the presence, absence, or price of the other items. In the absence of cross effects, the choice of item X is completely independent of any other item.

A disadvantage of SCE is that only significant cross-effects should be included and the significance of cross effects needs to be determined beforehand. In a situation with many items, it could become a tedious exercise that would benefit from automation. Fortunately, Sawtooth Software's MBC software provides a counts module that examines all potential cross-effects and points out which cross-effects seem to be potentially the strongest (assuming the underlying price design is efficient). However, we learned from our experience that it is not as a straightforward exercise to do as it might seem. Even though we used a balanced, orthogonal design, we found some significant cross effects according to counts that did not make sense. For a more comprehensive check it is advisable to run a series of 2-Log likelihood tests to decide which model structure suits the data best. Also, when the design is not balanced or orthogonal, the spotted cross-effects might just be artifacts from the design Some sort of "gut" feeling, market knowledge (art) and statistical expertise (science) is required for practitioners to build the right model. Apart from these challenges, another main drawback is related to the fact that, if cross effects are included (or not included), they hold (or are not present) for the entire sample. In case of heterogeneous samples, or much clustered samples, this might harm the predictive validity of the model when looking into subsamples.

EXHAUSTIVE ALTERNATIVES MODEL (EA)

An Exhaustive Alternatives model (EA) recognizes and estimates the combinatorial outcomes of all menu item choices in conjunction. This is in contrast with binary logit models, which estimate the marginal choices of each individual item. As such, EA represents a more comprehensive model of consumer choice. The model specification assumes that the agent makes a tradeoff between the choice of a combination of features against all the possible combinations. The dependent variable is the choice of a combination using a single logit-based model (MNL). The main drawback of this model is that the number of possible combinations grows exponentially with the number of items. This might become unfeasible with a large number of items—which is likely to happen in business applications. However, a solution could be to split the menu into portions or subgroups that can be coded up as Exhaustive Alternatives, thus leveraging many of the benefits of the exhaustive approach without leading to too many combinations to manage in practice.

Johnson, Orme, and Pinnell (2006) implemented an exhaustive aggregate logit model covering the full factorial set of alternatives, with the alternative configured by the respondent marked as "chosen," for each choice task. It is basically a main-effects model with only coding of effects for each attribute level. This approach produced nearly the same result when coding each attribute in a separate choice task, mimicking an "extreme partial-profile task" where all other attributes were held constant. It means that breaking up the task into a series of tasks encoding the information separately for each attribute leads to the same results as exhaustive coding. However, the authors used aggregate logit, so they did not get the benefits of HB estimation.

Later we present a different approach to tackle exhaustive alternative models. We use a mix of concepts borrowed from exhaustive alternative theory, traditional conjoint tool analysis, and HB estimation. First, we present our case study in the next section.

STUDY DESIGN

Although the data collected for this study were not extracted from a commercial study, we argue that the type of MBC study used in this paper is typical for MBC to be found in our practice. It was a study into consumer preferences for notebook computers (AKA laptops). The study contained twelve items: consumer features of notebook computers. A consumer feature is an added value to a notebook beyond the technical core functionality of the notebook. Each item or consumer feature came in three price levels; a base price and a higher or lower price. This was done to measure the consumer's willingness to pay for the consumer features. The price variations were specified in the absolute price change (a fixed amount up or down), not a relative price change. The prices varied across the items in accordance with an orthogonal research design. We also varied the core of the notebook, giving the respondent a choice among three technical specifications of a notebook, each at its own price point. An overview of the prices of the items is in the table below.

Table 1. Alternative price levels for the twelve features or "items" in our MBC study

	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6	Feature7	Feature8	Feature9	Feature 10	Feature 11	Feature 12
Price 1	$5.00	$15.00	$1.00	$1.00	$10.00	$15.00	$30.00	$30.00	$15.00	$15.00	$1.00	$3.00
Price 2	$10.00	$30.00	$3.00	$3.00	$20.00	$30.00	$50.00	$50.00	$25.00	$30.00	$3.00	$5.00
Price 3	$20.00	$50.00	$5.00	$5.00	$30.00	$50.00	$70.00	$70.00	$35.00	$50.00	$5.00	$7.00

The MBC exercise consisted of nine choice tasks; seven random choice tasks and two holdout tasks. The holdout tasks were placed randomly during the MBC exercise. The sample size was n= 1,408 consumers.

ANALYSIS PROCEDURE

To analyze the BYO exercise in our study we thought of remaining on the same line of reasoning of the exhaustive alternative models (but capturing the preferences as main effects for each attribute). We liked the idea of considering the tradeoff between combinations of items. As a matter of fact, the choice of a single item in the menu is still one of the possible combinations.

Let Ω be the choice set of all possible combinations that a respondent can select. The choice set Ω has 2^n combinations. In our study Ω accounts for a total of 3*2^12=12,888 possible combinations. This number of combinations would be computationally challenging and it would jeopardize the estimation. A solution, suggested by McFadden (1978) and Ben-Akiva and Lerman (1985), is by exploiting the independence from irrelevant alternatives (IIA) property of the logit models. It permits consistent estimation with only a subset of the combinations, which should include the chosen one and a sample of non-chosen combinations.

There are several methods of sampling alternatives with related model estimation procedures. One method is the *Random Sampling of Alternatives* where alternatives are randomly drawn from Ω. However this method is not efficient if the majority of alternatives have small choice probabilities. Therefore, we use a similar method, the *Importance Sampling of Alternatives*. It considers a sub-sample of combinations that have a higher probability to be chosen by the decision maker. We consider the choice set Θ ∈ Ω that only contains the chosen combinations. In our sample, out of the 12,888 possible combinations, “only” 4560 were chosen by respondents, indicating some relevance. For each respondent in each task we draw a random sample from this choice set Θ ∈ Ω .

Then we consider a multinomial logit where the probability of the respondent choosing a *k*th combination in each task is:

$$p_k = \frac{\exp\,(x_k^i \beta_i)}{\sum_i \exp\,(x_i^{'} \beta_i)}$$

Where :

$k \in \Theta$, the choice set of all the chosen combinations;
$p_k =$ the probability of individual *i* to choose a *k*th combination;
$x_i^{'}$= a vector describing the *j*th combination

In the exhaustive alternative models, $x_i^{'}$ is coded as a single attribute with *k* levels, each representing a combination and/or its price. i.e., in an MBC exercise with three dichotomous items, this attribute would have 2^3= 8 levels. Each item can have a single price attribute, or a total price attribute across all items. However, we code each item and its price in the $x_i^{'}$ matrix as separate attributes (instead of a unique attribute with all combinations as levels). Therefore, our model mimics a standard main-effect CBC task, and each task is indeed coded as described in the table below.

Table 2. Code frame of respondent choices in our MBC study

CASEID	Task#	Concept#	Core	Feature1	Price1	Feature 2	Price2	...	Response
1	1	**1**	1	1	2	1	3	...	**0**
1	1	**2**	1	2	0	1	3	...	**1**
1	1	**3**	1	2	0	2	0	...	**0**
...	...	**...**	...	...	...	...	...	...	**...**
1	1	**33**	2	2	0	1	3	...	**0**

The 33 concepts are the combinations drawn randomly from the choice set Θ —with one being the chosen combination. Each feature has two levels (Yes/No) and the corresponding price levels are alternative-specific according to the feature being in the combination or not. We basically translate a BYO task into a standard conjoint task.

Including SCE tasks

We include a further refinement to our model. BYO choice tasks are a gold mine of information at an individual level. We observe individual potential "price barriers" for each optional feature as we know whether a respondent chooses each feature at each price point, regardless of other items' price.

On top of the combinatorial model, we include a series of dummy tasks for each respondent and for each feature. Those are a series of binary logit tasks, where the choice of a feature is modeled as a function of its own price only (own effect), in isolation from the other effects.

We use the HB procedure to estimate individual utilities for each feature (Yes/No level), each feature price level (3 price level) and a None option. Resulting utilities are used to build a "what if" simulator both with Share of Preference (SoP) and Share of First choice (FC) method. The number of products in our simulator is the whole set of chosen combinations, Θ. In this way we are able to simulate both feature choices and combinatorial choices.

To sum up, the model is a "mixed approach": a combinatorial coding using a sampling of alternatives method, and a series of binary logit tasks. The latter mimics an "extreme partial profile task," in a similar fashion as in Johnson, Orme and Pinnell (2006).

However, the fact that we use HB estimation allows substitution effects to be revealed via simulations on idiosyncratic respondents using the logit rule, due to patterns of preference among pairs of items as reflected in HB heterogeneous preferences. Therefore, in the presence of substitution effects among the alternatives, the use of HB should cover most of the substitution behavior into our model and the resulting simulator. In the next section we empirically compare a CSS model with the SCE model as employed by the Sawtooth Software MBC module in the notebook computers study.

Comparing results

In our first attempt to test CSS models presented at the 2011 SKIM/Sawtooth Event, we also aimed at comparing CSS with SCE models. As previously explained, SCE require the identification of significant cross-effects to include in the models. We could not find significant cross-effects among our menu alternatives in terms of price cross-elasticity. We attributed this to the small sample size (N=200) and to the fact that the presented optional features did not have inter-independence in functionality. Therefore, we use a much bigger sample size in the present study (N=1,408) and we include features for which we expected a substitution effect, causing cross effects to occur.

Neither a count analysis with Chi-square tests nor a series of 2-Log Likelihood Tests to test different model specifications with aggregate logit reveal the presence of significant cross effects among alternatives. Therefore, when building the 12 different models for each feature choice for SCE, we only include the "own" price effects for each feature choice. Using Sawtooth Software's MBC, we estimate single feature choice predictions[27] and combinatorial choice predictions[28].

In order to compare results from the SCE and CSS models, we compare their internal validity in terms of hit rates of holdout tasks. The table below reports aggregate mean average error values (MAE) and R-square values for two holdout tasks.

Table 3. Aggregate mean average error values (MAE) and R-square values for two holdout tasks

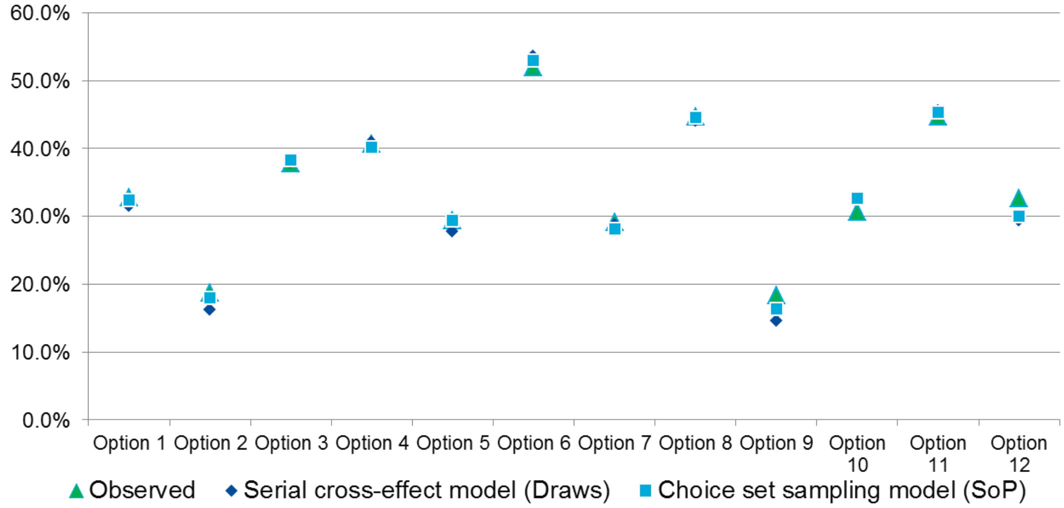
		Holdout task 1		Holdout task 2	
		R-Squared	MAE	R-Squared	MAE
Serial Cross Effect Model	HB, Point Estimates	0.991	0.90%	0.991	1.00%
	HB, Draws	0.991	0.90%	0.992	1.10%
Choice Set Sampling Model	HB, First Choice	0.987	1.60%	0.989	1.70%
	HB, Share of Preference	0.984	1.50%	0.981	1.50%

The accuracy of both models is pretty similar, with SCE slightly outperforming CSS. Comforting results arise from the error structure. We do not find any pattern in the errors across the models, meaning that we do not consistently overestimate or underestimate each single feature choice prediction. We deliberately included high and low priced features, and neither model is impacted by price magnitude and feature price difference. The chart below shows the error distribution for hold-out task 1. Both models are modeling observed choices with an average MAE that is never more than 5%.

[27] Using Draws and Point Estimates

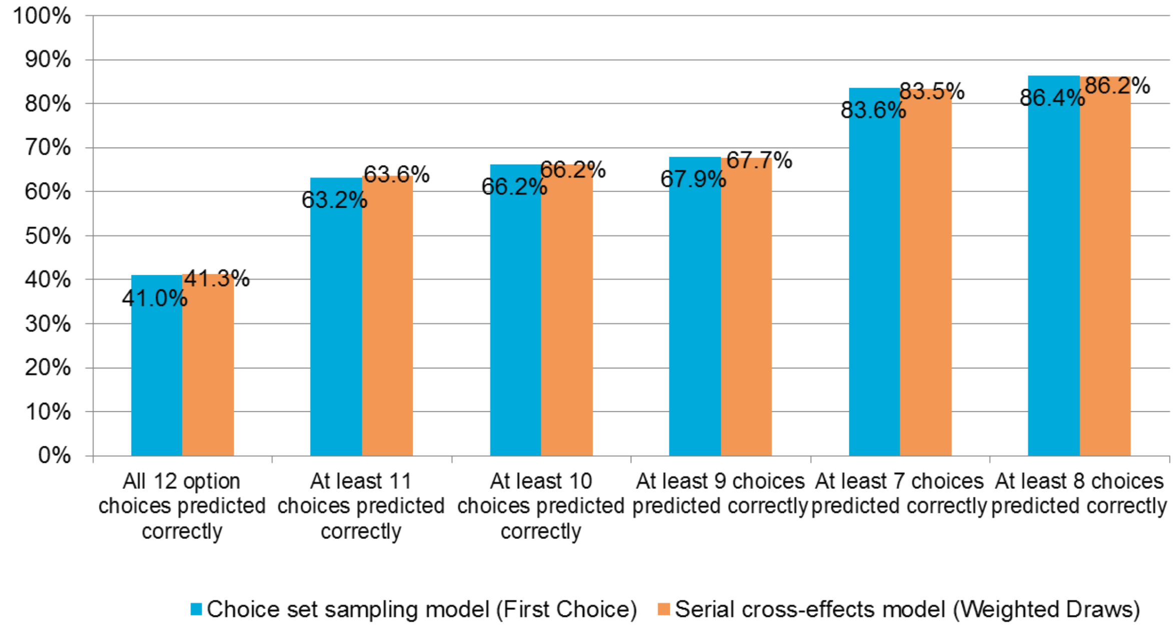
[28] Using Draws, Point Estimates and Weighted Draws.

Figure 1. Visual representation of the choices of menu items, observed versus predicted according to a SCE and CSS model



Another comforting result comes from the analysis of combinatorial predictions. The figure below shows the percentage of correct individual level combinatorial predictions for different number of features.

Figure 2. Visual representation of the shares of combinatorial choice of menu items, predicted according to a SCE and CSS model



With both models we correctly predict the chosen combination of 12 features for around 41% of the sample. Almost 10% of the sample did not choose any option. Considering all the possible combinations that could have been chosen (12,488), a hit rate of 41% is still quite a good prediction.

CONCLUSIONS AND FUTURE RESEARCH

Both models are effective to accurately predict holdout choice tasks at an aggregate level and individual-level choices of single options and combinations so they are both viable tools for analyzing MBC data. Now, as practitioners we are also interested in the efficiency when applying the models in a business context, where time pressure is a variable we cannot avoid. A Choice Set Sampling model (CSS) and its origin, an Exhaustive Alternatives model (EA), are more encapsulating approaches. The heterogeneity captured in the model (individual-level models) can account for substitutability of one alternative on the menu for the other, due to the fact the model accounts for substitution among combinatorial alternatives within the set. However, it is quite tedious to build them which might become overly time consuming. Another problem is that the resulting simulators are quite heavy in terms of their call upon computational power, which makes them inconvenient to clients.

On the other hand, SCE models are relatively quick to set up and estimate, especially now that Sawtooth Software has released its dedicated MBC software. The software delivers support for identifying relevant cross effects, as well as fast and easy simulation tools. Of course practitioners will have to go through a learning curve in understanding how to interpret the significance of cross effects and their inclusion in the model.

A caveat is that the user must be careful when including (or not including) cross-effects in the model, as they are applied to all respondents. This may render invalid when looking into subsamples of heterogeneous or clustered samples. Future research into this topic will need to address this. Ideally, we would use a SCE model that includes relevant cross effects at (semi-) individual level. A suggestion would be to include group-level cross effects detected by means of Latent Class analyses.

REFERENCES AND FURTHER READING

Bakken, D.G. and L.R. Bayer (2001), “Increasing the Value of Choice-based Conjoint with ‘Build Your Own’ Configuration Questions.” Sawtooth Software Conference Proceedings, pp. 229-238.

Bakken, D. and M. K. Bond, “Estimating Preferences for Product Bundles vs. *a la carte* Choices,” in Proceedings of the Sawtooth Software Conference, October 2004, Sequim, Washington: Sawtooth Software (2005).

Ben-Akiva, M. and Gershenfeld, S. (1998), “Multi-Featured Products and Services: Analyzing Pricing and Bundling Strategies”. Journal of Forecasting volume 17, Issue 3-4, page 175–196.

Chrzan, Keith (2004), “The Options Pricing Model: A Pricing Application of Best-Worst Measurement,” 2004 Sawtooth Software Conference Proceedings, Sequim, WA.

Horne, Jack, Silvo Lenart, Bob Rayner and Paul Donagher (2010), “An Empirical Test of Bundling Techniques for Choice Modeling,” Sawtooth Software Conference Proceedings, pp. 77-90.

Johnson, Richard M., Bryan Orme, and Jon Pinnell (2006). “Simulating Market Preference with ‘Build Your Own’ Data,” Sawtooth Software Conference Proceedings, pp. 239-253.

Liechty, John, Venkatram Ramaswamy and Steven H. Cohen (2001), “Choice Menus for Mass Customization: An experimental approach for analyzing customer demand with an application to a Web-based information service”. Journal of Marketing Research, 39 (2), 183-196.

Moore, Chris (2010), “Analyzing Pick n’ Mix Menus via Choice Analysis to Optimize the Client Portfolio,” Sawtooth Software Conference Proceedings, pp. 59-76.

Orme, Bryan K. (2010), “Menu-Based Choice Modeling Using Traditional Tools,” Sawtooth Software Conference Proceedings, pp. 85-110.

ABOUT THE AUTHORS

Carlo Borghi is a Project Manager at SKIM with a split responsibility in innovation and in conducting research projects in SKIM’s telecom, finance and durables business. Carlo has a Master’s degree in Mathematics from the University of Leiden in the Netherlands.

Paolo Cordella is a Research Consultant in SKIM’s Research Services & Software division. He developed the menu-based choice modeling application partly during his post-graduate internship at SKIM. Paolo has a Master’s degree in Economics from the Catholic University of Leuven in Belgium.

Kees van der Wagt is a Research Director at SKIM. He divides his time between overseeing and inspiring the Innovation in methodology projects, and running client projects for SKIM’s Research Services & Software division. Kees has a Master’s degree in Econometrics from Erasmus University in Rotterdam, the Netherlands.

Gerard Loosschilder is SKIM’s Chief Methodology Officer. He is responsible for SKIM’s strategy in methodology and market/client driven innovation. Gerard has a PhD in market research from the Delft University of Technology in Delft, The Netherlands.

BUILDING EXPANDABLE VOLUME CONSUMPTION ONTO A SHARE ONLY MNL MODEL

ROHIT PANDEY
COLUMBIA UNIVERSITY
JOHN WAGNER
ROBYN KNAPPENBERGER
NIELSEN

ABSTRACT

In this paper, we investigate methods to turn share only logit models into expandable volume models, where the total model volume can change depending upon the scenario defined [prices, availability, etc.]. Two general approaches were attempted: 1) regressing volume against full-worth utilities using respondent-level multinomial logit models estimated ahead of time with HB (Hierarchical Bayes) estimation; 2) including a "Lost Volume" (hence forth, LV) term into the logit models, where this additional term may be estimated jointly with the other parameters of the share model or estimated separately after the other terms.

We conclude that the respondent-level volumetric data has so much variability that the resulting regression-based models suffer from an unacceptable amount of volatility.

We found some success with the lost volume approach. With the HB models, we found that estimating the lost volume term separately works best. More recent development utilizes a smoothing function on the respondent data and Aggregate Logit with Sourcing (ALS), a model developed internally. We see this newest approach as one with great potential.

INTRODUCTION

The Multinomial Logit Model in Conjoint Analysis

Conjoint analysis is becoming a widely used tool in market research. The term conjoint comes from the English word, *conjoin* or *conjoined*, but marketers often think of it in terms of "respondents *con*sider the items *jointly*." It is assumed that the respondents, instead of seeing a product as a whole, consider an item to be a combination of features (like price) and associate utilities with each of these features, ultimately choosing the item that has the highest combined utility for them. The most common model used in choice-based conjoint analysis today is the multinomial logit model (a.k.a. MNL). This model is used for estimating the shares of various items based on certain combinations of attributes' levels. Each of these attribute levels has a parameter to estimate. The estimation can be done either at the aggregate level or (with new estimation techniques like HB) at the respondent level. Once the parameters of the multinomial logit model have been determined (through any of the various estimation techniques available) the researcher can determine the shares of all items based on a combination of these attributes' levels. For example, this model is often used in studies for soft drinks, where (in the most basic form) there are two attributes, brand and price.

Need for Expandable Volume

In FMCG (Fast Moving Consumer Good) categories, other quantities, most notably the volume sold, also contribute important information. Hence, there is a need for an expandable volume model, which can change the total volume as scenarios change. Using market data on the total volume consumed from the category and the results of the share only model, the item level volumes are captured. These in turn lead to other quantities of interest for the client (such as profits, revenue, etc.). This model works fine, except that being a share model, the volume for the category remains fixed (often to the current market scenario). This is a restrictive assumption, as we would expect the volume of the category to differ as we change the scenarios. For example, if some of the brands increase their prices, not only will their shares go down, but the more loyal users might start consuming less from the category (either switching to other categories or decreasing their usage). If the respondents are asked how many units they would buy in the choice tasks, this information on the movement of the category is present in the data, but cannot be captured by the share only multinomial logit model. Hence, there arises the need for a model that can capture the patterns of the movement of category volume and provide a more realistic estimate of the volumes of the items.

Literature Review

At the beginning of this work, we chose simple models, based on easy to implement techniques. Hence, the regression-based approaches given in Garrat and Eagle [1] and Méndez and Montenegro [2], based on conferences we (or other peers) had attended were the best place to start. A couple of potential approaches were described in [1], beginning from pure regression, using the choice model and the regression approach we picked which is described as Joint discrete-continuous (combining choice models with regression). In it, the denominator of the multinomial logit equation is assumed to be an indicator of how good a trip (a specific choice set among those a respondent sees) is for a respondent and volume is regressed linearly on the denominator to get coefficients generating scenario-wise variable volume. In Mendez & Montenegro 2010 [2], the authors describe a method for turning the preference share model (that multinomial logit is based on) into a volume share model, and then extending the resulting volume shares into actual respondent level volume using regression with the top three utilities. The appendices in the manual of the Sawtooth Software's CBC-HB show that a very similar procedure is internally used by Sawtooth Software to deal with chip allocations (multiple choices and weights for the items instead of one) instead of preference as the authors in [2] describe. From the Sawtooth Software conference (which was the source of [1]), we also went through the work of Bryan Orme [4] (which was the basis for a Sawtooth Software suggestion for modeling expandable volume). In it, the None term from conjoint is used for modeling volume movements. The term is however modified in the data file by taking the non-None maximum across respondent trips and subtracting from it the total volume in any trip to compute the None allocation for that trip.

For a study (referred to henceforth as study 1), various ideas for modeling this expandable category volume were tried. These could broadly be grouped into regression-based models and lost volume models, which are explained in the following sections.

Regression-based Approaches

These approaches attempt to add expandable volume to the study once the share model has been run. They tend to rely on some function of the utilities of the scenario that one would expect to be high for scenarios with attractive alternatives at low prices (where a respondent would choose high volume) and low for scenarios with high prices and/or unattractive alternatives (where a respondent would choose less volume).

A respondent would be expected to choose more volume on a trip when she sees choices she really likes. The denominator-based approaches to regression assume that this is reflected in the utilities of the items in the trips. So, it is logical to choose a function that is representative of the utilities of all items of the trip (or at least the most important ones to the respondent). One example of such a function is the denominator of the multinomial logit equation. This term is the sum across all alternatives of the exponentiated full-worth utilities of each alternative. Because of the high magnitudes of this term, we take the logarithm of this factor, instead of using it as is. The equation from [1] can be expressed as:

$$V_{scn} = B_0 + B_1.\log(D_{scn})$$

Where V_{scn} is the total volume is a given scenario "scn" and D_{scn} is the MNL denominator for that scenario. Note also that a scenario is something we simulate while a trip refers to a choice set seen by a respondent.

Another variant of this model that we tried was:

$$\log(V_{scn}) = B_0 + B_1.\log(D_{scn})$$

A variant of this approach assumes that the respondent doesn't care about all the items in the trip and only considers the ones that she would consider buying. Of course, those would be the items with the highest utilities. It is reasonable to assume that a respondent would, on an average trip, purchase three of her favorite items in some quantities. If we denote the top three utilities as U_{max}, U_{max-1} and $U_{max-2,}$ the regression equation becomes[29]:

$$V_{scn} = (B_0 + B_1.U_{max} + B_2.U_{max-1} + B_3.U_{max-2})$$

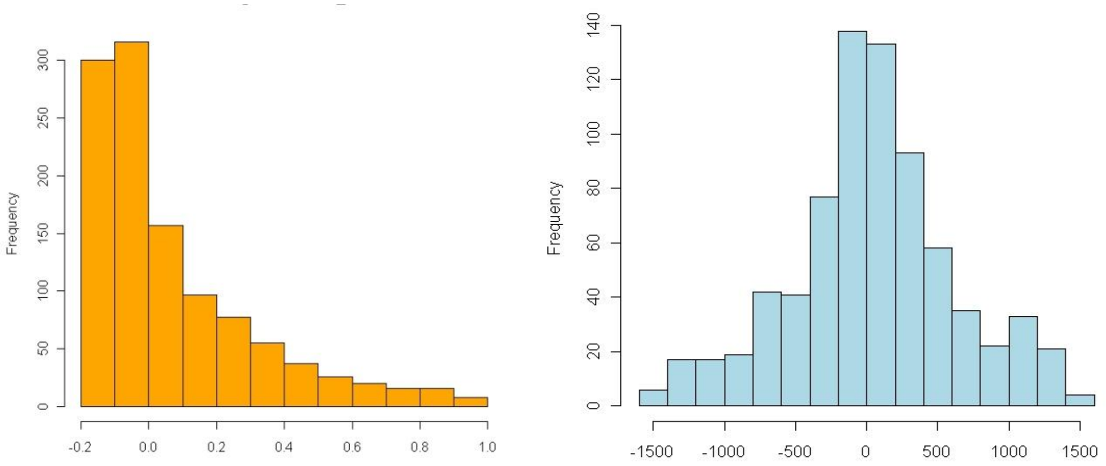
As before, we tried a variant of this approach:

$$\log(V_{scn}) = (B_0 + B_1.U_{max} + B_2.U_{max-1} + B_3.U_{max-2})$$

Where U_{max-i} is the $(i-1)^{th}$ largest item utility of all those a respondent sees and the B_i's are just coefficients.

[29] In the source document [2], a slightly different form is used which takes the logarithms of the utilities instead of taking them as is. We decided against this as utilities can potentially have negative values, which could render the equation unsolvable.

The regression based methods didn’t do too well as was clear from the actual values of the coefficients we got (you can see that close to half of them were the wrong sign in the histogram below) and also the adjusted R^2 values (shown in a histogram below).



Histograms of the adjusted R^2 values and regression coefficients across respondents

The following are features of this approach that make it undesirable:

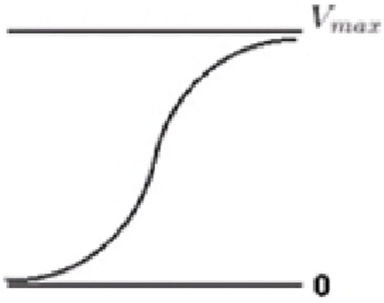
- Respondent data is typically very sparse and random. Hence, expecting each respondent to follow predefined expectations (those dictated by a simple linear model) is naive. This is reflected in the poor fits, many of which have negative adjusted R^2 values.
- Because we are meshing two models together (the share model and the expandable volume model) there is the possibility that inconsistencies can creep into the model. This phenomenon is best explained through an example. Suppose we increase the price of an item, thereby decreasing its utility. This will have the following effects:

  1. The share of that item would go down
  2. The shares of all the other items would go up
  3. The total category volume would come down

  Now, consider the effect of the second and third points on the predicted volume of an item unrelated to the price increase. These are two opposing factors and it might sometimes happen that the decrease in category volume supersedes the drop in share, meaning that the volume of this item actually decreases. This is counter intuitive, as increasing the price of one item should have no effect on the volume of an unrelated item. Our attempts with real data had these counterintuitive effects occur quite often.
- Since the model is linear, the predicted volume can range from $-\infty$ to $+\infty$. Yet negative volume doesn’t exist; and volume will have a limit on the positive end (even if the items were given away for free, there is a limit as to how much product would be carried out of the store). Hence, a good volumetric model would be:

1. Bounded below at zero (hence asymptotic to it) and bounded above (and asymptotic) at the maximum volume a respondent can be expected to purchase. This upper bound will be represented henceforth by V_{max}.
2. Monotonically increasing
3. Smooth and continuous everywhere.

If we try to keep all these properties in mind and draw a curve, the most obvious shape one can imagine is an S shaped curve. This is shown in the figure below:



An S-shaped curve

Lost Volume Approaches

Our second wave of attempts at modeling volume involves estimating a “lost volume” term for the respondents. Lost volume indicates the volume that the respondent could potentially have bought, but didn't because she didn't like the trip as much as she could have. Hence, this term is interpreted as the lost volume potential (described in the literature review and [4]). In common market research choice modeling parlance, the lost volume might be called the ‘share of none’. The advantages of this approach are:

- Since there is just one model, there is no scope for inconsistencies as defined in the previous section.
- As explained in the previous section, an expandable volume model which by its nature is monotonically increasing and bounded will follow an S-shaped curve. This is an inherent feature of the lost volume method. This is because it builds the movement seamlessly into the multinomial logit equation, for which any item's share follows an S-shaped curve.

Because the lost volume approach uses the multinomial logit model, the combined share of the items (excluding the lost volume term), follows an S-shaped curve with respect to the utilities of any of the items (or a group of them). So, it is asymptotic at zero volume and at some upper bound. This upper bound would be the maximum volume we would expect the respondent to buy in any scenario. The natural choices for the upper bound could be:

- The maximum volume a respondent chooses across the choice trips (and if all of them are random, meaning none is particularly good, we could consider scaling this maximum up by some factor).

- A value for maximum volume volunteered by a respondent as a result of a direct question.
- The volume chosen by a respondent in a particularly “good” trip (in this case, the design would need to accommodate this. For example, a particularly good trip - like all items available at their lowest prices - could be built into the design).

In general, the size of the lost volume term has the following effects on the outputs of the multinomial logit model:

- We should expect the shares of the expandable volume model to not deviate too much from the “share only” model (which was what we started with – no movement of category volume). In practice, we observed that the larger the “lost volume” term, the more the deviation in the item shares between the two models.
- A large lost volume term can cause large volume movements. If the lost volume term is large, it tends to draw (or forfeit) large shares as scenarios change, causing larger movement in category volume.

In the following sections, we investigate various approaches based on this methodology.

The first one was based on estimating the lost volume term as a parameter (henceforth, pre-estimation LV). This approach for calculating the lost volume term is mentioned in [4]. Here, the term is estimated just like all the other parameters, with the allocation for this term replaced by the maximum choice across trips minus the total volume in that trip. Some observations from estimating the lost volume term through the HB methodology were:

- It is clear that estimating a lost volume term has a detrimental effect on the Root Likelihood (RLH) values, meaning poorer model fit. So, higher values for such terms lead to lower RLH's in general.
- The lost volume term is generally too large, leading to excessively large volume movements as scenarios change.

The next was based on estimating the lost volume term post estimation (henceforth, post estimation LV). Since our pre-estimation approach resulted in a very large lost volume term, we experimented with an approach where we estimate a share model, fix the parameters for the share model, then add in a lost volume term and determine its value in a second step. We will explain this now in further detail.

Let’s assume that the MNL utility for the LV alternative is L, for a given respondent (while the other terms are as defined above). The equation that relates the lost volume term to the volume chosen in a trip is:

$$\frac{D_{scn}}{D_{scn}+e^{L}}=\frac{V_{scn}}{V_{\max}} \qquad (1)$$

The intuitive value of equation (1) is that as L → ∞, meaning that the lost volume term is taking up all the volume, $V_{scn} \rightarrow 0$. And, as L → −∞, the L.H.S of the equation starts tending to 1 and so, $V_{scn} \rightarrow V_{max}$.

Since this equation has two unknowns (V_{scn} and L), it is impossible to solve for both of them. We need to calculate L and for that, we want one more piece of information. The trips we have from the respondents could serve this purpose. We could feed the equation the information we have on any trip's volume and logit denominator. Substituting this information into the above equation would mean that it produces the actual volume for the trip we used. So, this choice is like defining a base for the model. If we have 5 choice tasks for a respondent, then the next question is which one do we take? Of course, this choice would affect the value of L and consequently, the respondent-level volumes at any scenario. Whichever we choose, we would get predictions for the volumes of the rest of the four choice tasks for a respondent, and also have the actual volumes chosen in those tasks. We can try the different trips as the choice for the one to be used in calculating L, find the corresponding volumes for the trip we already have for the respondent and see which of these produces the closest results. Let's say this establishes which trip to use and label it t. The denominator for that trip is D_t and the volume for it is V_t. Using equation (1) we get,

$$\frac{D_t}{D_t + e^L} = \frac{V_t}{V_{\max}}$$

$$\Rightarrow e^L = D_t\left(\frac{V_{\max}}{V_t} - 1\right)$$

$$\Rightarrow L = \log\left[D_t\left(\frac{V_{\max}}{V_t} - 1\right)\right]$$

Substituting this in equation (1) gives us:

$$\frac{V_{scn}}{V_{\max}} = \frac{D_{scn}}{D_{scn} + D_t\left(\frac{V_{\max}}{V_t} - 1\right)}$$

$$\Rightarrow V_{scn} = V_{\max}\cdot\left(\frac{D_{scn}}{D_{scn} + D_t\left(\frac{V_{\max} - V_t}{V_t}\right)}\right)$$

This is the equation we would use to get V_t. So, in a way we have five models that can be expressed as a matrix with each column representing a different model (with each of the trips as input). Here, $v_{i,j}$ denotes the prediction of item j's volume by model i. Note that the i^{th} model would produce the actual trip volume (V_i) for the i^{th} item and so the matrix would look something like:

$$\begin{bmatrix} V_1 & v_{1,2} & v_{1,3} & v_{1,4} & v_{1,5} \\ v_{2,1} & V_2 & v_{2,3} & v_{2,4} & v_{2,5} \\ v_{3,1} & v_{3,2} & V_3 & v_{3,4} & v_{3,5} \\ v_{4,1} & v_{4,2} & v_{4,3} & V_4 & v_{4,5} \\ v_{5,1} & v_{5,2} & v_{5,3} & v_{5,4} & V_5 \end{bmatrix}$$

Among the five sets represented by each of the columns of this matrix, we see which one matches closest to the actual trip volumes (V_1 to V_5). The criterion used can be the mean absolute error (MAE) or the covariance. Tests have shown the MAE approach to produce the most satisfactory results.

The above approach for estimating the LV term post estimation is easy to implement (which is of particular value as we have to do this for each respondent). A more natural way to go about this would be to pick the L described above from a continuous set (like -∞ to +∞) as opposed to picking it from a discrete pool based on the trips of a respondent. In order to do this, we should think of L as a continuous term. Hence, we could form an expression for the MAE (or a related expression) in terms of L and find the value of L that minimizes this expression.

Suppose the volume predicted by the model for a given scenario is V_{scn}. Suppose also that this scenario happens to be a trip of that respondent, where she chose volume V_t. The squared error term associated with this prediction would then be:

$$e_t = (V_{scn} - V_t)^2$$

From equation (1), this becomes:

$$e_t = \left(V_{\max} \frac{D_t}{D_t + e^L} - V_t \right)^2$$

And the sum of these errors across the trips (denoted here by E) would become:

$$E = \sum_t e_t = \sum_t \left(V_{\max} \frac{D_t}{D_t + e^L} - V_t \right)^2$$

We employ numerical methods to optimize this function.

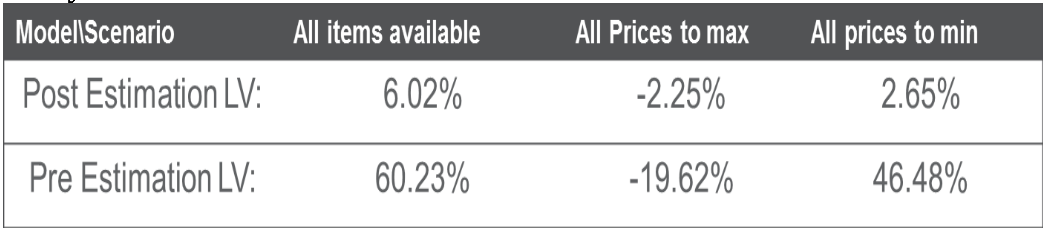
RESULTS AND CONCLUSIONS

As mentioned before, the regression based approaches did not fare well, with counter intuitive signs for the coefficients, very poor fits and the presence of inconsistencies (described before), leading to illogical behavior (like improving the characteristics of one item causing the shares of another to increase).

The lost volume approaches fared much better than the regression approaches. They have all the properties associated with what we would expect from a good expandable consumption model. Our initial work showed that the post-estimation lost volume approach led to more reasonable outcomes than pre-estimation lost volume (where the entire estimation is carried out with the HB approach). We present below, some empirical results to this end:

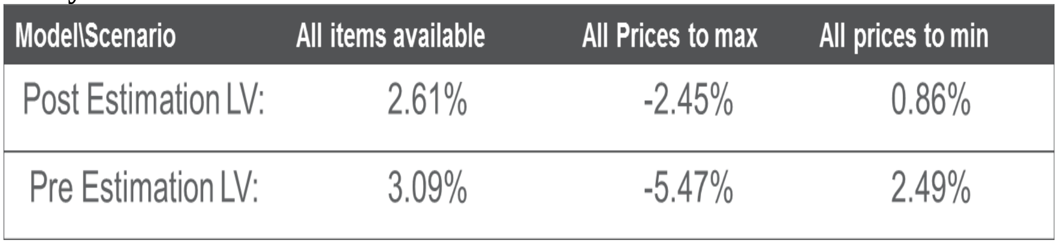
The first two tables compare the category volume movements from the two studies we tested:

Study 1:



Model\Scenario	All items available	All Prices to max	All prices to min
Post Estimation LV:	6.02%	-2.25%	2.65%
Pre Estimation LV:	60.23%	-19.62%	46.48%

Study 2:

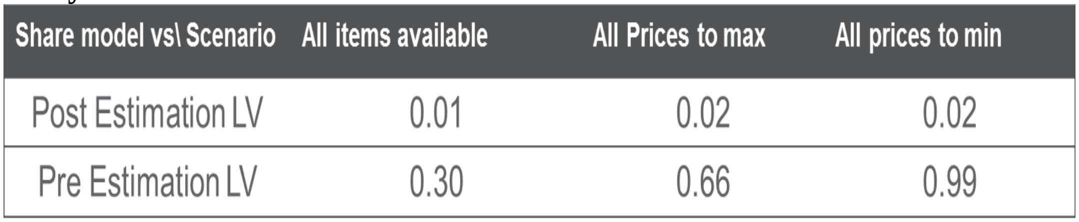


Model\Scenario	All items available	All Prices to max	All prices to min
Post Estimation LV:	2.61%	-2.45%	0.86%
Pre Estimation LV:	3.09%	-5.47%	2.49%

As can be seen, the volume movements are far less volatile with the post estimation lost volume model and this is more in line with what we would expect in practice.

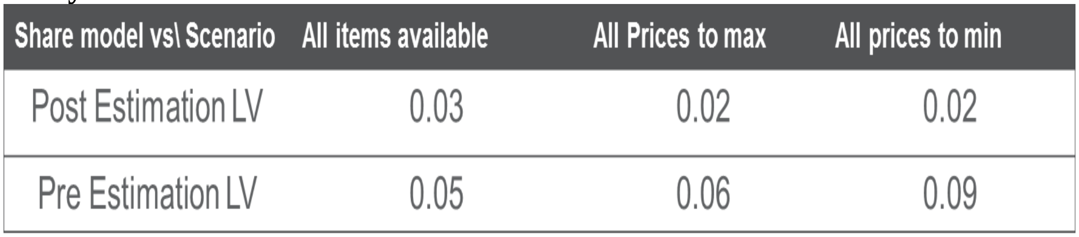
The tables below provide a comparison of the alignment (based on the Mean Absolute Error) of the two models with the share only model (total volume across scenarios remains the same). We would want this alignment to be good as the share model is well-tested with a lot of validation backing it up.

Study 1:



Share model vs\ Scenario	All items available	All Prices to max	All prices to min
Post Estimation LV	0.01	0.02	0.02
Pre Estimation LV	0.30	0.66	0.99

Study 2:



Share model vs\ Scenario	All items available	All Prices to max	All prices to min
Post Estimation LV	0.03	0.02	0.02
Pre Estimation LV	0.05	0.06	0.09

Again, the post estimation lost volume model aligns better with the share only model for the corresponding scenarios. Hence, we can conclude that estimating the lost volume term post estimation leads to better results (than doing it along with the other parameters – referred to as pre estimation lost volume) when using HB.

In subsequent sections, we provide a model that takes the approach of applying the pre estimation lost volume estimation methodology to a different model

ENHANCEMENTS TO THE LOST VOLUME APPROACHES

After taking some time away from this development, we picked it up with new ideas to build on our previous research. We have implemented two fundamental enhancements to the pre-estimation lost volume approach and have since successfully implemented these on live studies for clients. There are still some limitations to this approach because of the inherent weaknesses in the respondent data we use [respondent data quality variability, stockpiling versus increased consumption, etc.], but our initial results are promising.

The two enhancements are the following:

1. Smoothing of the Respondent Data,
2. Use of a new share model, the Aggregate Logit with Sourcing model [ALS].

We will now describe in detail these two enhancements.

Smoothing of the Respondent Data

Our approach creates a function that takes as its input the list of total volumes chosen by a respondent across her tasks. The function maps those total raw volumes to what we call modeled volumes. We have two pictorial ways to describe how the map works using an example.

In our example, the raw total volumes per task chosen by the respondent are 45, 40, 35, 30, 25, 20, 20, 14, 12, 10, 8, and 4, from greatest to least [the order of tasks is not relevant for this approach]. These total volumes per task get mapped respectively to 25, 24, 23, 22, 21, 20, 20, 17, 16, 15, 14, and 12. The modeled volumes are whole numbers only because we artificially chose values for the raw total volumes that would lead to this result. Normally, the values would not round so nicely.

The table on the left shows the result of the smoothing map; the explanation on the right indicates how the map works.

Raw	Modeled
45	**25**
40	24
35	23
30	22
25	21
20	**20**
20	**20**
14	17
12	16
10	15
8	14
4	**12**

Step 1: the median [20] is mapped to itself [20].

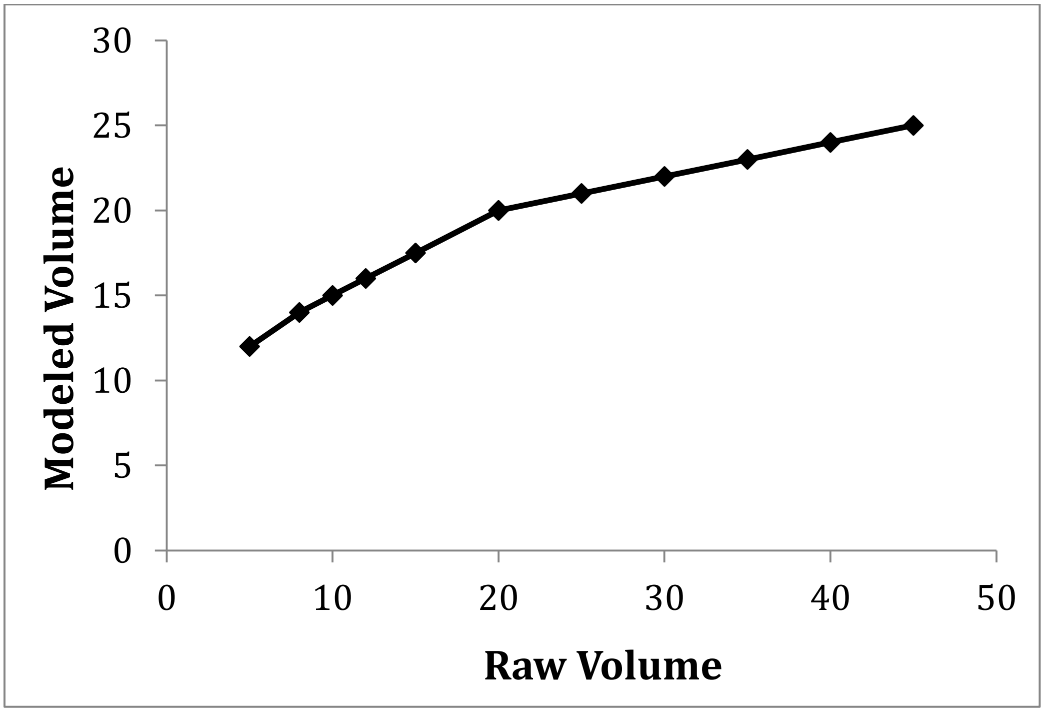
Step 2: the greatest value [45] is mapped to the 5th greatest value [25].

Step 3: values between the greatest and the median are interpolated linearly.

Step 4: the least value [4] is mapped to the 4th least value [12].

Step 5: values between the least and the median are interpolated linearly.

The graph of this example map is the following:



There are two important aspects of this approach to note about the choices of the 5th greatest value and the 4th least value of the raw data as the respective values for the modeled maximum and minimum.

1. These choices are somewhat arbitrary. The smoothing is determined by applying outside influences and influenced by the modeler, not determined by the data itself. We'd prefer to have better data. Given the data we currently get, we feel this approach is reasonable and enables us to make a useful volumetric model.

2. These choices are asymmetric. We believe that the overstatement of above-average volume is worse than understatement of below-average volume. Stockpiling is the consumer behavior that is confounded with increased consumption in our survey data. Therefore, we decrease high volumes more than we increase low volumes.

Depending on the category of the study, and sometimes depending upon the results we get with the model, we can make different choices [i.e. instead of taking the 5th greatest raw value and mapping the raw maximum to it, we may take the 3rd greatest value; and likewise for the 4th least value].

For each task, each of the raw data choices is multiplied by the ratio of the modeled volume to the raw volume. This multiplication results in the total raw volume for a task to be correctly mapped to the total modeled volume for the task according to this map we have described.

The share relationship across the items in a given task is maintained, and the volatility of the respondent data is decreased significantly.

Aggregate Logit with Sourcing (ALS)

[30]ALS is a proprietary model (Nielsen patent pending) we developed to overcome the problem that occurs when the property of Independence from Irrelevant Alternatives (IIA) inherent in the multinomial logit model is not appropriate for the study at hand. Some of our studies include products which are clearly not equally substitutable [e.g. colas, diet colas, caffeine-free colas, etc.]. Prior to developing ALS, we have used extensively an HB MNL model, which has the property of IIA on the respondent level. The HB MNL model is certainly appropriate for many studies, but in some cases IIA on the respondent level is inappropriate.

In our studies we have used one item attribute with a number of levels equal to the number of products included in the study. In addition to the item attribute, typically we have only a price attribute, which we convert to ln(current price / base price), and treat as a linear attribute. Occasionally we have additional promotional attributes which get treated in a similar manner as the price attribute. We will leave the details of those additional promotional attributes out of this discussion. We additionally have a sourcing attribute which, working together with MNL submodels, enables ALS to overcome the property of IIA. The number of submodels is equal to the number of items in the study. Here are the details.

Each level of the item attribute has an associated item utility parameter, u_i. This is very similar to the item utility parameter you would have using an aggregate multinomial logit model.

The sourcing parameters are structured in a square matrix. The number of rows and columns in the matrix is equal to the number of items in the study. Each row corresponds

[30] ALS is proprietary (Nielsen patent pending). (Editorial Note: The conference steering committee was not aware that the methods presented in this paper would be claimed as proprietary and patent pending. It is generally not appropriate to present proprietary work at the Sawtooth Software Conference that cannot be fully disclosed and offered to the community as industry knowledge. We will strive to make it more clear in future calls for papers regarding our desire that the conference be an open forum for presenting and discussing methodologies.)

to a submodel, and each column corresponds to an item. We denote the element of the matrix in row r, column c as $m_{r,c}$.

The intercept for an item in a submodel is a sum of the item's utility parameter, u_i, and the element of the sourcing matrix in the row corresponding to the submodel and the column corresponding to the item. The resulting formula for the intercept term for item i in submodel s is $u_i + m_{s,i}$.

We place restrictions on the matrix of sourcing parameters. Firstly, the diagonal elements are fixed at zero. Secondly, the off-diagonal elements are forced to be negative. These two restrictions are important to ALS. Consider the intercepts for item #1 in a model where there are four items, and thus four submodels:

$u_1 + m_{1,1}$ $u_1 + m_{2,1}$ $u_1 + m_{3,1}$ $u_1 + m_{4,1}$	Since $m_{1,1}$ is zero and each of $m_{2,1}$, $m_{3,1}$, and $m_{4,1}$ are negative, the intercept for this item is maximized in the submodel #1. The same phenomenon occurs with item #2: its intercept is maximized in submodel #2. And so on, for the other items. By these means, we can say that each item is 'featured' in a specific submodel which is associated with it.

The weights for the submodels further enable this 'featuring' effect. The item utility parameters are used as the means to determine the submodel weights. The calculation is the same calculation as for choice probability of the multinomial logit model:

$w_1 = \exp(u_1) \; / \; [\exp(u_1) + \exp(u_2) + \ldots + \exp(u_4)]$
where exp (x) means e^x.

The connection between each item and its associated submodel enables an intuitive understanding of how ALS[31] handles sourcing. Consider that if all of the sourcing parameters were zero, then each submodel would merely be a duplicate of the same multinomial logit model, with intercepts $u_1, u_2, \ldots, u_4$. This model would have fair-share sourcing (i.e. IIA). Now consider the first submodel, which uses the first row of the matrix of sourcing parameters. As the sourcing parameters in this row move in the negative direction, those which move farther from zero force the sourcing of the corresponding item to be lower than those whose sourcing parameters stay closer to zero. So if $m_{1,2} = -3$ and $m_{1,3} = -0.5$, item #1 will source more from item #3 than from #2, as compared to fair share.

An additional restriction which we sometimes employ on our ALS[32] models is to force the matrix of sourcing parameters to be symmetric. This approach cuts in half the number of sourcing parameters yet still seems to enable the model to capture the sourcing patterns revealed via the respondent data.

Let's look at an example. Consider a simple setup with four items.

[31] ALS is proprietary (Nielsen patent pending).
[32] ALS is proprietary (Nielsen patent pending).

		Item 1	Item 2	Item 3	Item 4
	item utilities:	0	1	0	1
matrix of	submodel 1:	0	-0.5	-3	-3
sourcing	submodel 2:	-0.5	0	-3	-3
parameters	submodel 3:	-3	-3	0	-0.5
	submodel 4:	-3	-3	-0.5	0

The intercepts for the submodels are the following:

						Submodel
		Item 1	Item 2	Item 3	Item 4	Weight
Intercepts	submodel 1	0	0.5	-3	-2	13.4%
for the	submodel 2	-0.5	1	-3	-2	36.6%
sub models	submodel 3	-3	-2	0	0.5	13.4%
	submodel 4	-3	-2	-0.5	1	36.6%

The weights for the submodels are calculated as described above using the item utilities. Once the model is estimated, the weights remain fixed for the simulation of any scenario.

The choice probabilities at base conditions are the following:

						Submodel
		Item 1	Item 2	Item 3	Item 4	Weight
choice	submodel 1	35.3%	58.2%	1.8%	4.8%	13.4%
probabilities	submodel 2	17.3%	77.4%	1.4%	3.9%	36.6%
	submodel 3	1.8%	4.8%	35.3%	58.2%	13.4%
	submodel 4	1.4%	3.9%	17.3%	77.4%	36.6%
	Weight-averaged	11.8%	38.2%	11.8%	38.2%	

You can see that both items #2 and #4 have the same share at base conditions [no effect of the prices]. Now let’s consider what happens when the item utility for item #1 is decreased [presumably by an increase in price]. Let the item utility for item #1 decrease from 0 to -1. Here are the resulting choice probabilities, per submodel and for the model’s total:

		Item 1	Item 2	Item 3	Item 4	Submodel Weight
choice probabilities	submodel 1	16.7%	74.9%	2.3%	6.1%	13.4%
	submodel 2	7.1%	86.9%	1.6%	4.3%	36.6%
	submodel 3	0.7%	4.8%	35.7%	58.8%	13.4%
	submodel 4	0.5%	3.9%	17.4%	78.1%	36.6%
	Weight-averaged	5.1%	43.9%	12.1%	38.9%	

Here is a table of comparison that shows clearly the fact that item #1 sources much more from item #2 than either of item #3 or item #4.

	Item 1	Item 2	Item 3	Item 4
Base Conditions	11.8%	38.2%	11.8%	38.2%
Item1 at high price	5.1%	43.9%	12.1%	38.9%
Difference	-6.7%	5.7%	0.2%	0.7%

Our current approach to capture response to price uses both a price per item parameter and a price per submodel parameter. They get summed together respectively: the price slope term for item #1 in submodel #3 is equal to the sum of the price parameter for item #1 and the price parameter for submodel #3. This approach has proven satisfactory across a wide range of studies. It enables sufficient flexibility in capturing complex price responses in the respondent data while not requiring an excessive number of model parameters.

ALS[33], like the HB MNL, is a discrete mixed-logit model, consisting of multinomial logit submodels which produce choice probabilities for all of the items which get aggregated to produce the full model's choice shares. The HB MNL model identifies each submodel with an individual respondent, and in estimation calculates likelihoods based on how well that submodel predicts the respondent's choices. In this respect, ALS is quite different from HB MNL. With ALS, the likelihood calculation is based on how well the full model predicts the choices in each of the tasks.

Summary and Conclusions from our Enhancement Work

Both the smoothing of the respondent data and the new Aggregate Sourcing with Logit (ALS) model are working quite well for us. The smoothing function enables us to model data which comes to us raw with less-than-credible volatility regarding volume. For clients who absolutely require variable volume modeling, we believe that this is the best way to go

There are still limitations, however, because we're taking far-from-perfect data and applying somewhat arbitrary choices to make that data fit what seems to be reasonable. We know that the data inherently possesses issues with stockpiling vs. increased

[33] ALS is proprietary (Nielsen patent pending).

consumption and with the conversion of non-category buyers into category buyers. In the future we hope to incorporate much better data into the process not only with better choice data but also perhaps by incorporating retail data.

ALS, on the other hand, seems to hold great promise as an all-around share model. It successfully captures sourcing patterns which do not follow IIA [Independence from Irrelevant Alternatives]. It uses a closed form calculation of choice probabilities, which is quite fast and thus in this respect far superior to probit-based models. It is also free from assumptions on the distribution of variables across respondents, which affect Hierarchical Bayes models [such as CBC/HB]. Our work with respondent-level volumetric data shows the inherent volatility and leads us to believe that an aggregate model has an advantage over respondent-level models for volumetric modeling. Furthermore, since ALS does not have the IIA property on the respondent level, it is more suitable than HB models for handling categories where individual respondents can show patterns of sourcing which do not follow IIA.

REFERENCES

Mark Garratt, Thomas C. Eagle; Practical approaches to modeling demand (Based on 2010 Sawtooth Software conference)

Miguel Teixidor Méndez, Julián Sánchez Montenegro; Continuous Choice Analysis (Based on 2010 SKIM conference, Barcelona)

Kenneth.E.Train; Discrete choice methods with simulation (Text book)

Bryan Orme; Menu-Based Choice Modeling Using Traditional Tools (Based on 2010 Sawtooth Software conference)

CREATING TARGETED INITIAL POPULATIONS FOR GENETIC PRODUCT SEARCHES

SCOTT FERGUSON
CALLAWAY TURNER
GARRETT FOSTER
NORTH CAROLINA STATE UNIVERSITY

JOSEPH DONNDELINGER
MARK BELTRAMO
GENERAL MOTORS RESEARCH AND DEVELOPMENT

ABSTRACT

Genetic searches often use randomly generated initial populations that maximize genetic diversity and thoroughly sample the design space. While many of the initial configurations perform poorly, the tradeoff between population diversity and solution quality is typically acceptable for small design spaces. However, as the design space grows in complexity, computational savings and initial solution quality become more significant considerations. This paper explores the benefits associated with the strategic construction of a “targeted” initial population using respondent-level utilities from a discrete choice model. Two case study problems are presented to illustrate the benefits of this approach, including reductions in computational cost, improvements in the preference shares of product search solutions, and more consistent performance of product search solutions on problem objectives. Additionally, this work extends creation of the targeted initial population to a multiobjective problem formulation; this provides for exploration of trade-offs between competing business objectives, a capability not currently supported in Sawtooth Software’s SMRT.

1. INTRODUCTION

Product designers are often faced with the challenge of creating products for markets with highly heterogeneous customer preferences. Doing this in a competitive – and profitable – manner requires balancing product variety with the costs of product development and production. Striking this balance becomes particularly difficult when solving large-scale combinatorial content offering problems. To offer some perspective, consider the scale of automotive feature packaging problems: the number of available features can be in the hundreds and, in practice, the numbers of product variants can be hundreds or even thousands (even though the scale of our optimization problem is only dozens). When faced with a problem of this magnitude, it becomes readily apparent that solving for optimal product configurations using exhaustive search algorithms is intractable. Supporting this conclusion is the fact that one billion design evaluations, assuming each one takes one second, would require 31 years to complete. This number is significant, as our second case study problem has 1,074,954,240 possible build combinations to consider.

Advancements in optimization have yielded multiple techniques capable of providing solutions to problems that are intractable by an exhaustive search [1, 2]. Often, these algorithms are tailored to solve problems of specific form: 1) single variable / multivariate, 2) linear / nonlinear, 3) unconstrained / constrained, and 4) single objective / multiobjective. To speed convergence toward a single solution, many of these approaches rely on information from the objective function to calculate gradients (first-order information) and the Hessian (second-order information), requiring analytical derivatives or estimation using numerical methods.

Feature packaging problems, or problems of similar form, pose a significant challenge for many optimization techniques. Product features typically cannot be represented in a continuous design space. Instead, the levels in a study may represent a feature's presence / absence, or the various means by which that feature can be implemented. As shown in Table 1, there is often no intelligible meaning for a solution that exists between integer levels. For example, assigning a value of 1.5 to *Attribute 1* implies that the attribute is both on and off. Since this is not possible, gradient-based optimization techniques cannot effectively be used, as the gradient is calculated with the assumption of a continuous design space. Further, since some attributes (such as price) can be represented as a continuous variable, a product designer is faced with finding optimal solutions to a mixed-integer problem.

Table 1. Attributes and associated levels

Attribute level	Attribute 1	Attribute 2	Attribute 3
1	On	Low	$50
2	Off	Medium	$100
3		High	$150

Sawtooth Software's Advanced Simulation Module (ASM) offers a variety of product search algorithms [3]. For problems where gradient information cannot be used, ASM converts the gradient and stochastic searches to a grid search. A grid search is completed by using a heuristic process to randomly select a product attribute and examine all permitted levels of that attribute. The best level of that attribute is retained and the process repeats until the process fails to return a better solution. While significantly faster to converge than an exhaustive search, a grid search is not guaranteed to find the global optimum in a multi-modal space.

Genetic search is a particularly attractive option for these problems, as it naturally handles discrete information, is zero-order, and can navigate multi-modal performance spaces. Introduced nearly 30 years ago by John Holland [4], the genetic search uses a population-based search strategy. Individual product line solutions are represented by a string of values (similar to a DNA sequence) and a series of these strings represents the current population of solutions. Techniques called crossover and mutation mimic the reproduction of a DNA string, creating a new set of solutions to evaluate. At each generation, the best strings are retained and the process repeats until the algorithm reaches some stopping criteria.

The mixed-integer nature of feature packaging problems, combined with the possible combinatorial complexity as more attributes are considered, suggests that a genetic search might be the most effective optimization technique. However, as shown in Figure 1, results obtained from ASM for an automobile feature packaging dataset show that the grid search outperforms the genetic search. As both search techniques are stochastic in nature, each optimization was completed five times. The error bars depict the maximum and minimum preference share found using each search technique, and the marker represents the mean.

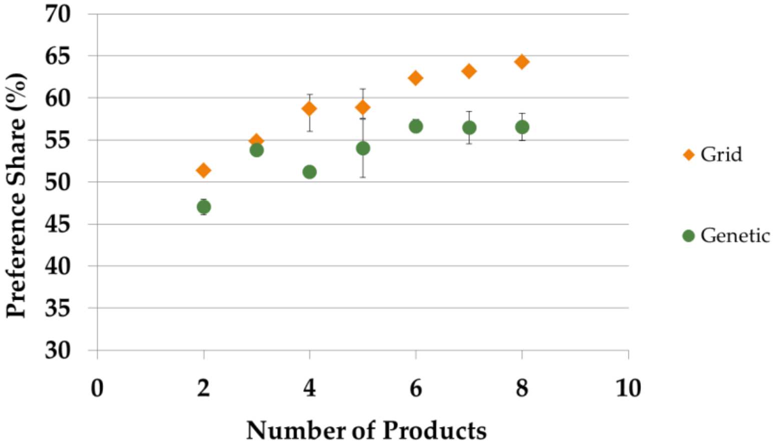


Figure 1. Comparison of gradient and genetic search performance in ASM

Of significance in Figure 1 is that share of preference does not monotonically increase for the genetic search as additional products are included in the market simulation. This outcome strongly suggests that the genetic search is either terminating at a local optimum or is being limited by the genetic information in the initial population of products. While the mutation operator in a genetic search provides a means of escaping from local optimum, methods for generating effective initial populations have been severely understudied. Thus the objective of this paper is to present an approach for improving the starting population of the genetic search so that the desirable regions of the design space can be more efficiently located and explored.

The next section of this paper provides a background discussion of product line optimization and fundamental properties of genetic searches. An approach to designing effective starting populations is described in Section 3 and results of two case study problems are presented in Section 4. Section 5 extends this approach to multiobjective optimization problems and generation of structured market solutions. Finally, conclusions and avenues of future research are discussed in Section 6.

2. Background

This section provides a brief background on optimization of product and product line designs from both the marketing research and engineering design literature. For problems with discrete variables, genetic search has been found to be an effective optimization strategy. A background on the operation of the genetic search is then presented with emphasis on the implications of applying a randomly generated initial population on the efficiency of the genetic search.

2.1 Product and product line optimization

Research in market-based product design has become increasingly popular over the last 20 years as the availability of computational resources has substantially increased. Optimization techniques that extend beyond traditional analytical approaches have been developed to handle complex design spaces. Michalek et al. proposed an extension of analytical target cascading (ATC) [5] to optimize the design of a single product when the market is represented by a multinomial logit model. This formulation is efficient at coordinating information flow between the marketing and engineering domains; however, it is not suitable for products with discrete attributes. An increasing amount of effort has been focused on applying heuristic techniques to conduct the optimization. Camm et al., for example, used branch-and-bound to determine an optimal single product [6]. Balakrishnan and Jacob used a genetic algorithm [7], while Besharati et al. expanded on this concept by using a multiobjective genetic algorithm to explore the tradeoff between predicted share and share variation due to uncertainty [8, 9].

For product line optimization, Green and Krieger implemented a greedy heuristic [10] and other heuristic rules were summarily developed [11-15]. More analytical methods have been used by McBride and Zufryden in the form of linear programming [16], while Hanson and Martin explored traditional avenues of optimizing a product line with multinomial logit [17]. In this vein, Michalek et al. extended their ATC work in [5] to product line optimization [18].

Heuristic techniques have also received significant attention in product line optimization. Wang et al. modified the branch-and-bound approach in [6] to product line optimization by reformulating it as a branch-and-price method [19]. This two-level approach uses configuration design at the upper level and price optimization at the lower level. Genetic algorithms have also seen increased use [20-22]. In a study of 12 different optimization methods, Belloni et al. demonstrated that heuristic techniques – such as branch-and-bound, genetic algorithms, and simulated annealing – were more effective than greedy algorithms and analytical approaches like Lagrangian relaxation [23]. Significantly, they also claim that genetic algorithms are more easily set up than many other heuristic approaches. Tsararakis et al. expand on these studies by using the newer particle swarm optimization technique to grow the number of population-based algorithms in product line design [24].

Most significant to this work are the results presented by Balakrishnan et al. [25]. Using a genetic algorithm, they studied the influences of genetic search parameters on solution quality. In addition to assessing the impacts of novel crossover and mutation strategies, they also seeded their initial population with results from a Dynamic Programming heuristic. The arbitrary nature of the Dynamic Programming sequencing algorithm, however, allows for good solutions to be discarded early in the process.

In this paper, we develop the idea of intelligently seeding a genetic search using part-worth utilities from individual respondents. While this approach is discussed in Section 3, a further discussion on the implications of a random initial population in a genetic search is presented in the next section.

2.2 Implications of a random initial population

Research focused on improving the performance of genetic searches in engineering optimization has traditionally explored how different crossover [26] and mutation [27] techniques can increase computational efficiency or maximize solution diversity. Efforts addressing initial population creation often advocate the use of random draws, with research leading to algorithms that ensure diversity amongst the initial designs [28]. For mathematical test problems and most engineering optimization scenarios, a diverse initial population that completely samples the design space is a desirable characteristic to have. While many of these initial designs may perform poorly when evaluated by the problem's objective function, the unknown structure of the final solution requires genetic diversity to enable a proper search. This is especially attractive when the number of design variables in the problem is small, as the computational expense required to solve the problem is low.

As the number of design variables increases, computational expense of performing the genetic search grows exponentially, thus reducing the number of required function calls becomes a priority. But what happens when a population is randomly generated? To demonstrate this, consider the hypothetical product attributes shown in Table 2.

Table 2. Hypothetical attributes for initial population creation

Attribute level	Body style	Engine	Wheel type
1	Sports car	V6 gasoline	Sport
2	Pickup	V8 gasoline	Off-road
3	Sedan	8 cylinder diesel	Passenger

Creating an initial population using random draws can lead to acceptable starting designs – as well as designs that are undesirable if not absurd. Figure 2 is an example of what happens when a reasonable design is created by merging a pickup body style, 8 cylinder diesel engine, and passenger tires. Now, consider another randomly created design where a sports car body style is merged with a V8 gasoline engine and off-road tires. As shown in Figure 3, the result is a technically feasible design that has a very low probability of marketplace success.

Figure 2. Example of a "good" randomly generated design

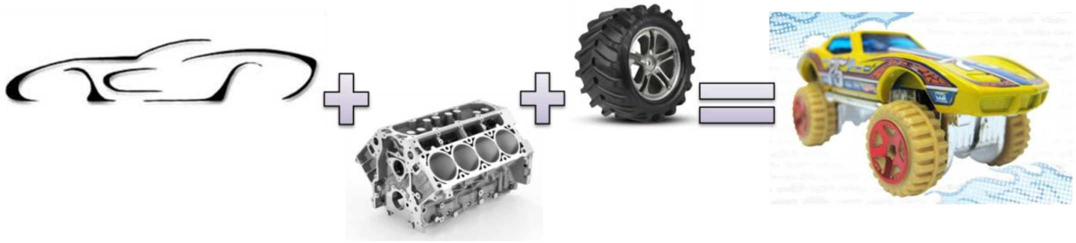


Figure 3. Example of a "bad" randomly generated design

Proceeding in this manner is akin to blindly throwing darts at a dart board. While many of the designs may hit the board, very few designs will be located near the bulls-eye (the optimal design in this analogy). This strategy has come to be accepted because there is rarely insight into how the starting population may be tailored by including product characteristics known to be desirable.

However, when integrating market research with engineering product design, a customer's part-worth utilities for each attribute level are known. These part-worth utilities give product designers a rich source of information that can be used to identify designs preferred by individuals, as shown in Figure 4. An approach for applying this information to speed up convergence and increase genetic search effectiveness is presented in the next section.

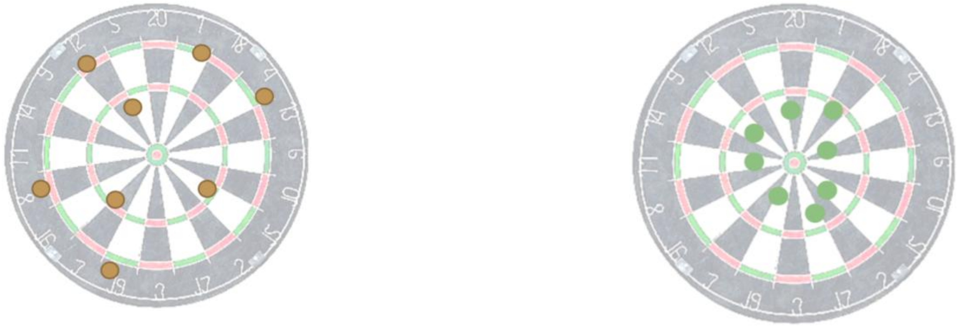


Figure 4. Random starting location (left) versus intelligently created starting designs (right)

3. Creating a targeted initial population

Homogeneous choice models, such as the multinomial logit, mask the heterogeneity of preference that might be present in a population of respondents. This is done by calculating a single set of part-worth utilities across all respondents for each attribute level. When heterogeneous models are used in product design, a respondent's utility for an alternative is represented by decomposing it into an observable component (V_{nj}) and an unobservable component (ε_{nj}) as in Equation 1.

$$U_{nj} = V_{nj} + \varepsilon_{nj} \qquad j = 1 \ldots J \tag{1}$$

The observable component of utility (V_{nj}) - or representative utility - is itself expressed as a function of alternative attributes (x_{nj}), coefficients of the alternative attributes (β), and consumer attributes (s_n), as in Equation 2. Alternative attributes encompass performance, price, and system characteristics. The observable utility is a function of the choice scenarios encountered and are estimated statistically.

$$V_{nj} = f(x_{nj}, \beta, s_n) \tag{2}$$

The approach presented in this paper is intended for problems where the market is heterogeneous and part-worth utilities have been captured at the respondent level. To obtain this information, respondents complete a choice-based conjoint battery generated by Sawtooth Software's CBC Web [29] tool. Part-worths for each individual are then estimated using Sawtooth Software's CBC/HB [30] module. While we use the results from CBC/HB in this work, the approach presented in this section is generalizable to any method capable of determining utility functions for each individual.

For a set of product offerings, the consumer part-worths are used to calculate share of preference. This information can then be aggregated with respect to the specified performance objectives. Section 3.1 describes how product solutions are created for the individual respondents, and Section 3.2 discusses how the individual product solutions are aggregated to create a product line for use in the initial population.

3.1 Finding optimized products for each respondent

The first step of this approach is finding the optimal product for each respondent. The intent is to create initial product lines by drawing from a pool of products that are optimal for at least one respondent. It is not guaranteed that each optimal product will perform well at the market level or capture new pockets of market share in a product line scenario. However, it is reasonable to expect that if at least one respondent finds a product highly desirable, it could serve as an effective starting point for the genetic search.

Optimal products are found for each respondent by performing a series of sub-optimizations to identify the attribute level settings that maximize observable utility. The formulation of this optimization sub-problem is shown in Equation 3.

$$Max: \; V_{nj} = f\left(x_{nj}, \beta, s_n\right) \tag{3}$$

$$with\ respect\ to: x_{nj},\ price\ of\ including\ x_{nj}$$

To perform these sub-optimizations, it is also necessary to assign price structures for each attribute; otherwise no tradeoffs would be necessary and the ideal product would simply consist of the most preferred level of every attribute at the lowest surveyed price point. An example price structure is shown in Table 3.

Table 3. Price structure for a hypothetical product attribute

Attribute level	Attribute setting	Attribute price
1	Off	$0 ('none' selected)
2	Low	$15
3	High	$50

Sub-optimizations are typically completed very quickly (often in seconds) and the cost of the optimization is relatively small because the size of the design space is greatly reduced. The objective function for each sub-optimization is finding the vector combination of respondent-level part-worths that give the maximum observable utility.

The computational cost of these sub-optimizations cannot be ignored as this would give this approach an unfair advantage when compared to other genetic searches that use a randomly generated initial population. Also, the number of function calls required to meet the defined stopping criteria must be considered in addition to the quality of the solution when comparing algorithms because, unlike runtime, the number of required objective function calls is independent of computer-specific attributes such as amount of memory available, processor speed, software used, etc.

Even so, objective function calls of the sub-optimization are not directly comparable to those of the product line optimization because the objective function is significantly less complex and the overall search space is reduced (only one product and one respondent are considered). Equation 4 is used to relate the sub-optimization space to the product line optimization space. In this equation, n is the number of respondents included in the study.

$$(Product\ Line\ Objective\ Function\ Calls) = \frac{1}{n}(Sub\text{-}optimization\ Function\ Calls) \quad (4)$$

The rationale for this relationship is that for each sub-optimization, respondent-level part-worths are used to determine the utility of a single product. However, for the product line optimization, utilities are calculated for multiple products for all respondents, and individual-level preference shares must be aggregated for the share of preference calculations. Based on run-time simulations (shown in Section 4.1) and the fact that the aggregation cost is not considered, Equation 4 is meant to be a conservative estimate of the true computational savings.

Having created an optimal design for each respondent, the next step of our approach is drawing from this pool to create initial product line offerings. This is discussed in the next section.

3.2 Generating initial product lines

The pool of ideal products provides source material for the initial product lines. While the number of ideal products is theoretically limited by the number of survey respondents (*n*), the number of initial product lines is an input parameter to the genetic search. Literature exploring effective input parameters to genetic searches suggests an initial population size approximately 7-10 times the number of design variables being considered [31].

As shown in Figure 5, the ideal products created in the previous task are now randomly selected, without replacements, and combined to create product lines. We consider this set of initial product lines to be a *targeted initial population* for the genetic search as it extends beyond the use of purely random configurations. Further, while heuristics exist to guide the number of initial product lines created, the number of products to be offered is a significant area of current and future research [21, 32].

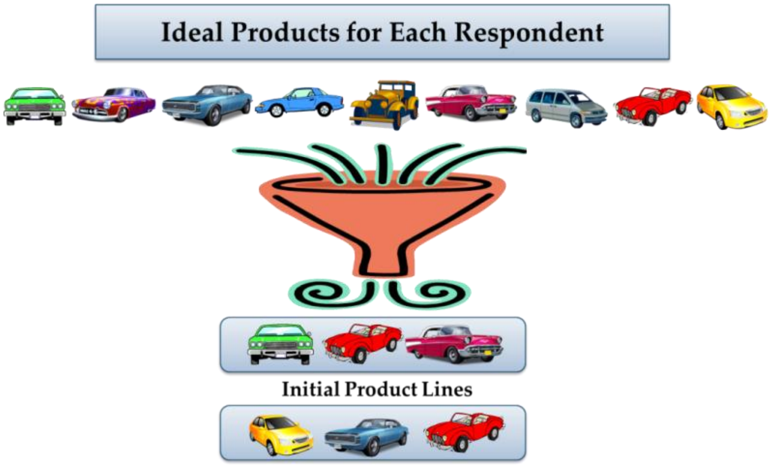


Figure 5. Illustration of creating targeted initial product lines

Generating a targeted initial population addresses the significant input for a genetic search. The remainder of the optimization may now be conducted using the operators of selection, crossover, and mutation. The next section discusses the initialization parameters used for our genetic search algorithm and the genetic search in Sawtooth Software's SMRT package.

3.3 Initialization parameters for the product line optimization

Key initialization advantages of the targeted initial population approach are the abilities to: 1) generate the initial population outside of the genetic search code, 2) generate the initial population in a parallel manner and 3) deploy the targeted population cross-platform. The only requirement for this approach is that the software package chosen for the genetic search allow the user to input / specify the initial population.

To assess the effectiveness of this approach, the following software scenarios were compared:

1) A genetic search conducted in Sawtooth Software's SMRT ASM package

2) A genetic search run in a Matlab environment with a randomly initialized population
3) A genetic search run in a Matlab environment with a targeted initial population

For the genetic search in SMRT (Case 1), the input parameters used are shown in Table 4. Also, for product line sizes greater than 5, software constraints limited *pool size* to a maximum of 1000 and the *offspring created within a generation* is limited to 500. The mutation rate was set at the default as defined within SMRT of 0.5.

Table 4. Input parameters for genetic search in SMRT

Criteria	Setting
Pool size	min(200*(size of product line),1000)
Offspring created within a generation	min(100*(size of product line),500)
Relative mutation rate	0.5
Stop after	20 generations without improvement
Report top	1 product

For the genetic search in Matlab [33] (Cases 2 and 3), the input parameters used are shown in Table 5. While the genetic search algorithms within the *Global Optimization Toolbox* (formally the *Genetic Algorithm and Direct Search Toolbox*) could be applied, a self-developed genetic search was used in this investigation.

Table 5. Input parameters for genetic search in a Matlab environment

Criteria	Setting
Pool size	min(100*(size of product line),500)
Offspring created within a generation	min(100*(size of product line),500)
Selection	Random
Crossover type	Uniform
Crossover rate	0.5
Mutation type	Gaussian
Mutation rate	0.5
Stop after	20 generations without improvement

Where possible, the standard GA operators were held constant between the genetic search in SMRT and the genetic searches in the Matlab environment. The only change in input parameters between Cases 2 and 3 is the specification of the initial population. Two Matlab-environment genetic searches are included to provide a comparison that removes algorithm bias.

Having described the approach for generating the targeted initial population and outlined the input parameters for the product line optimization, the next section of this paper reports simulation results from two case study problems.

4. Assessment of targeted population effectiveness

This section describes the effectiveness of applying a targeted initial population to two case study problems when using a genetic search. The case study problems differ significantly in size and results are presented from various stages in the product line optimization. In validating the effectiveness of the targeted population, it is expected that the final solution will:

- reduce the number of function evaluations required to meet the stopping criteria
- provide monotonically increasing preference shares as more products are introduced
- where possible, generate solutions that capture greater share of preference

As previously discussed, results will be shown from Sawtooth Software's SMRT genetic search, a Matlab-environment genetic search with a randomly generated initial population, and a Matlab-environment genetic search with a targeted initial population.

4.1 Case study 1: An MP3 player

The first case study problem presented in this paper is the result of a choice-based conjoint fielded by researchers at NC State and the University at Buffalo. The conjoint was created using Sawtooth Software's SSI Web; 140 students answered 10 questions each. As shown in Table 6, 10 product attributes were studied yielding a total of 23,040 possible feature combinations.

Table 6. MP3 player attributes

Photo Playback	Video Playback	Web Access	App Capable	Pedometer /Nike Support	Input Type	Display	Storage	Color	Price
Yes	Yes	Yes	Yes	Yes	Dial	1.5 in	2 GB Flash	Black	$76
No	No	No	No	No	Touch Pad	2 in	8 GB Flash	Silver	$149
					Touch Screen	2.5 in	16 GB Flash	Blue	$179
						3 in	32 GB Flash	Green	$229
						3.5 in	64 GB Flash	Red	$249
							160 GB HD	Orange	$299
								Pink	$399
								Custom	

This problem's design space is much smaller than in the automotive feature packaging problem that originally motivated this work, containing less than half the number of design variables; thus the initial pool size (initial population) was set to 200 for all product line sizes while the offspring within each generation was set at 100. The pricing structure for each product attribute is shown in Table 7. In addition to this pricing structure, a base price of $49 is added.

Table 7. MP3 player pricing structure

Photo Playback	Video Playback	Web Access	App Capable	Pedometer /Nike Support	Input Type	Display	Storage	Color
$5	$5	$20	$20	$10	$0	$27	$0	$0
$0	$0	$0	$0	$0	$15	$43	$18	$0
					$40	$58	$48	$10
						$69	$88	$10
						$77	$188	$10
							$85	$10
								$10
								$25

Creating the targeted population first required sub-optimizations to find each respondent's optimal product. Completing these sub-optimizations required an average of 233,940 total objective function calls from the genetic search. Applying Equation 4, this corresponds to 1671 objective function calls for the product line objective function.

Equation 4 was verified by comparing the time to complete 10,000 repetitions of four different simulation scenarios, as shown in Table 8. Comparison of these scenarios shows the increase in computational expense when evaluating increased numbers of both products and respondents. To ensure direct comparability of the results, all simulations were executed on the same computer. It was observed that 125 one respondent-one product calculations can be completed in the time necessary to complete one 140 respondent-one product calculation. Likewise, nearly 750 one respondent-one product calculations can be completed in the time required to conduct one 140 respondent-five product calculation. Altogether, this data confirms that Equation 4 provides conservative estimates of computational savings in initial population generation.

Table 8. Simulation time for utility calculation

Number of respondents	Number of products	Simulation time (seconds)
1	1	3.1e-5
1	5	1.6e-4
140	1	3.9e-3
140	5	2.4e-2

Genetic search was applied in this step because it was readily available and easily modified to match the problem formulation. However, a genetic search is not required for this step and is likely not the most efficient technique that could be applied. Since the objective is to maximize a product's utility in a multinomial logit model, many other techniques could be used.

Once the optimal products were found at the respondent level, the next step was to combine them into candidate product lines to form the targeted initial populations. In this case study, the product line was incremented from two to eight products. For each product line size scenario, 100 unique product line configurations were formed by randomly drawing from the pool of optimal products. Figure 6 shows the market-level shares of preference for each of these product line configurations.

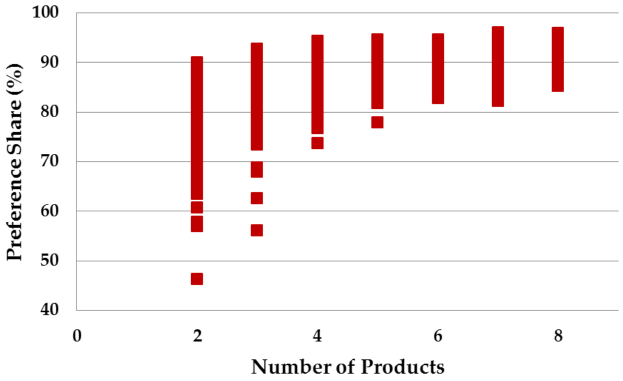


Figure 6. Share of preference distribution for the targeted initial population

The results of Figure 6 show that for each product line size, many product line configurations in the targeted initial population capture very large amounts of preference share. This trend seems to reveal an upper bound on potential preference shares for product line solutions. This is an intriguing result; it could suggest that for simple problems (those with small design spaces) one possible strategy is to generate a targeted initial population and simply select the best performing product line configuration within it. Further, these show that as the size of the product line increases, the preference share distribution tightens.

Comparing these results to preference shares of randomly generated initial populations reveals the true advantage of this approach. Figure 7 shows differences in average preference shares for targeted populations and randomly generated populations of equal size (100 product line configurations each). To ensure robustness of the reported result, 100 random initial and targeted populations were generated. Hence, the markers in the figure represent averages of average preference shares for the product line configurations within each initial population. The error bars on each marker indicate one standard deviation from the mean. They are not visible on the targeted population markers due to the scale of the figure; the average standard deviation for the targeted population was 0.42% while the average standard deviation for the randomly generated population was 1.69%.

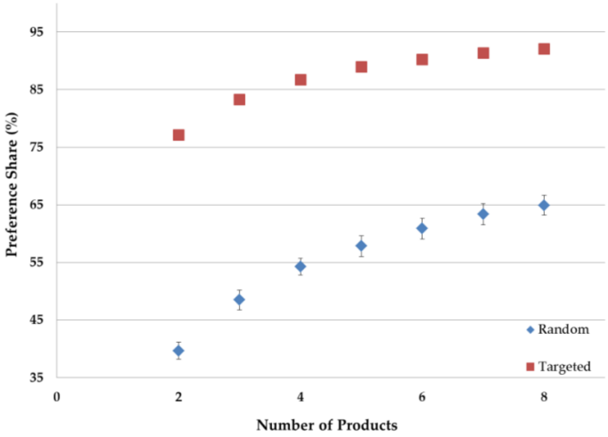


Figure 7. Performance comparison of initial population strategies

The results in Figure 7 do not offer a fair comparison because they do not account for the objective function calls incurred in generating the targeted population. To provide a more direct comparison, a genetic search is initiated using the randomly generated population and the genetic search operators listed in Table 5 and executed for 1671 objective function calls – the equivalent of the runtime required to generate the targeted population. A comparison of the maximum preference share solutions for five targeted and random populations when considered with equivalent runtime is shown in Figure 8.

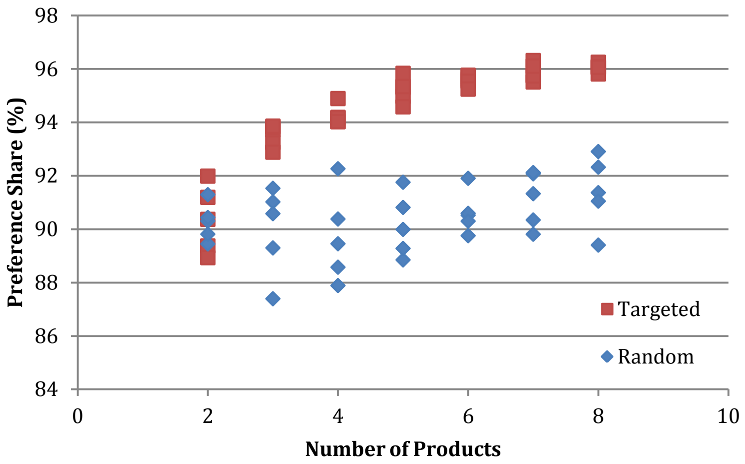


Figure 8. Performance comparison at equivalent objective function evaluations

Some interesting conclusions may be drawn from the results in Figure 8. First, when small product lines are created within a small design space, the randomly generated initial population is capable of achieving an equivalent intermediate solution. This is because for problems of this size, the randomly generated population may perform several generations of genetic operations within the runtime required to create the targeted population. However, when the complexity of the product line increases (causing the genetic string representing a product line solution to elongate), the best random population solution consistently underperforms compared to the best targeted population solution. The added length of the genetic string increases the difficulty for the crossover and mutation operators to achieve significant performance gains in a limited number of objective function evaluations.

Having compared the results from a random and targeted population, how does the targeted initial population solution compare to the solution found by the Sawtooth Software GA? Figure 9 shows the maximum preference share solution identified by the two different approaches. However, while the Sawtooth Software GA has been run until the stopping criterion from Table 5 was satisfied, the targeted initial population solution has just been generated. Figure 10 depicts the performance of the two approaches when both are allowed to run until their respective stopping criteria have been satisfied.

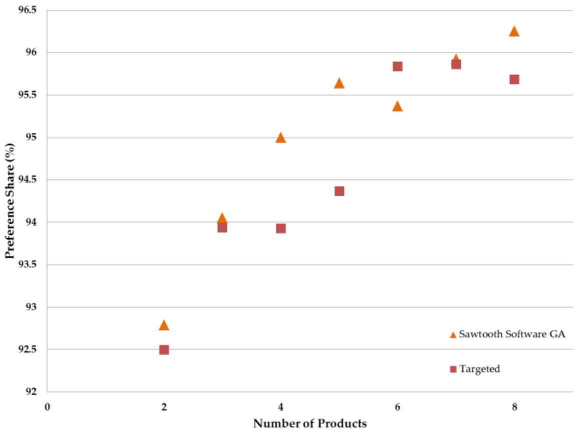


Figure 9. First comparison of targeted population and Sawtooth Software GA

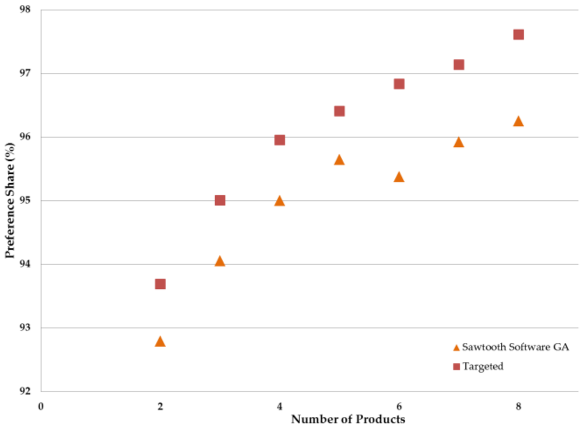


Figure 10. Comparison of targeted population and Sawtooth Software GA when both are run to completion

The results in Figure 10 show consistent improvements. First, the best product line solution achieved by the targeted population consistently outperforms the best product line solution found by the Sawtooth Software GA. Additionally, there is a monotonic increase in preference share as the size of the product line increases; this trend is not exhibited in the solution found by the Sawtooth Software GA.

While the targeted population obtains a better overall solution, computational cost (measured in terms of objective function calls) must also be compared. The number of evaluations required by the Sawtooth Software GA is shown in Table 9. Two types of information are reported in Table 10. The first is the number of function calls required by the targeted population to match the best solution obtained by the Sawtooth Software GA. The second piece of information is the number of function evaluations required to satisfy the stopping criterion.

Table 9. Objective function evaluations needed for the Sawtooth Software GA to satisfy stopping criterion

Number of products	2	3	4	5	6	7	8
Number of evaluations	2800	3400	3500	3400	4200	3600	3900

Table 10. Objective function evaluations needed for the targeted initial population genetic search in a Matlab environment

Number of products	2	3	4	5	6	7	8
Evaluations to match	2291	2161	2880	3297	1771	2771	3261
Total evaluations	2991	3861	5880	7997	7070	9471	9561

Comparing the numbers of objective function calls reported in Tables 9 and 10 shows that the targeted population approach can match the best solution from the Sawtooth Software GA with lower computational cost. It is more difficult, however, to draw conclusions regarding the total number of evaluations used by the two approaches. While the targeted population requires more functional calls to satisfy the stopping criterion, it achieves a better final product line solution. Additionally, these two approaches are coded in different pieces of software. Although the parameters of the genetic operators were matched as closely as possible, there may be inherent differences in the codes that prevent a direct comparison.

It was previously shown that the targeted population outperformed the random population. Figure 11 shows the best performance obtained by the random population genetic search in the Matlab environment and the Sawtooth Software GA when using the same number of objective function evaluations. This shows that the random population genetic search is capable of finding similar, if not better, solutions than the Sawtooth Software GA for the same computational cost. Therefore, we argue that the random population can be used as a surrogate for the Sawtooth Software GA when exploring performance at completion. Comparing genetic searches executed in the Matlab environment with targeted and random populations is more appropriate; except for definition of the initial population, all other aspects of the genetic search are identical.

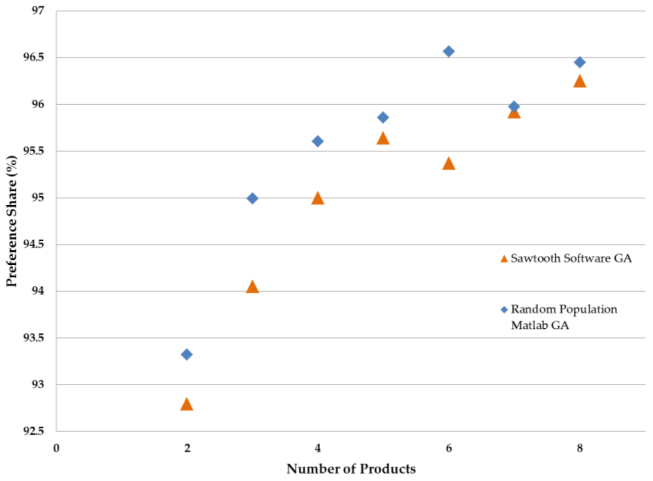


Figure 11. Comparison of random population GA and Sawtooth Software GA

When the targeted population genetic search and the random population genetic search are run to completion in Matlab, there are no appreciable differences in solutions for smaller product line sizes. As shown in Figure 12, it is only when the product line size exceeds 5 or more products that the targeted population approach is able to achieve a better solution. However, even when solution quality is the same, the targeted population typically achieves the final solution with a lower computational cost. This information is shown in Table 11.

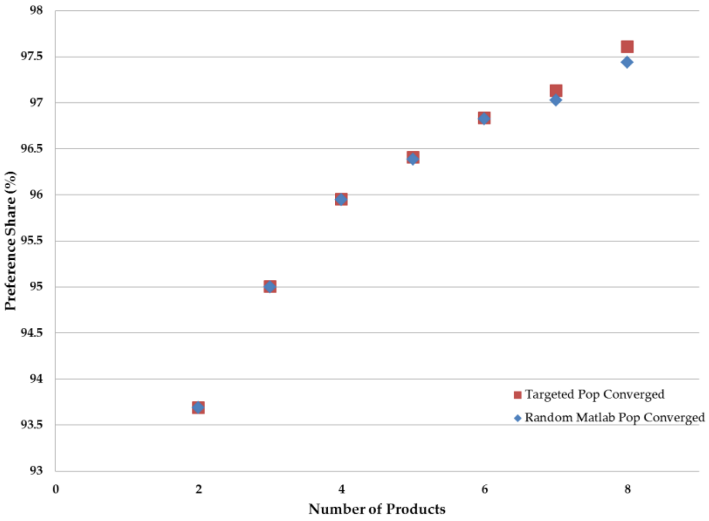


Figure 12. Comparison of random population GA and targeted population GA at completion

Table 11. Total number of objective function evaluations needed for the random population GA and the targeted population GA

Number of products	2	3	4	5	6	7	8
Random population	6400	4900	7500	7700	10700	8500	11200
Targeted population	2991	3861	5880	7997	7070	9471	9561

Having demonstrated the benefit of a targeted initial population on a problem with a relatively small design space, the next section of this paper explores a second, larger case study problem.

4.2 Case study 2: Automobile feature packaging problem

The second case study problem - an automobile feature packaging problem - originally motivated this research because of the complexity associated with the product line optimization problem. Results for this case study are based on a choice-based conjoint study commissioned by General Motors, in which 19 choice task questions were answered by 2275 respondents. Nineteen product attributes were considered in this study; Table 12 shows the number of levels associated with each attribute. Enumeration of all the possible feature combinations yields over 1,074,954,240 possible product configurations that must be considered.

Table 12. Levels per attribute for the automobile feature packaging problem

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}
3	2	5	6	2	3	3	2	4	2	3	2	4	3	3	4	4	3	2

Pricing structure information similar to that shown in Table 7 was used in this work; however, because of its proprietary nature, it cannot be disclosed in this paper. As in the first case study problem, the first step was to determine the optimal product solution at the respondent-level. These products were then randomly combined to create initial product line solutions. For this problem, the number of unique product line configurations created for the initial population was set at 100 times the product line size as this is approximately the same as 10 times the number of design variables [31]. However, for a product line containing 5 products or more, the initial population was capped at 1000 as this is the maximum allowable pool size within Sawtooth Software's SMRT. The market-level preference shares were estimated for each product line in the targeted population; the results are shown in Figure 13.

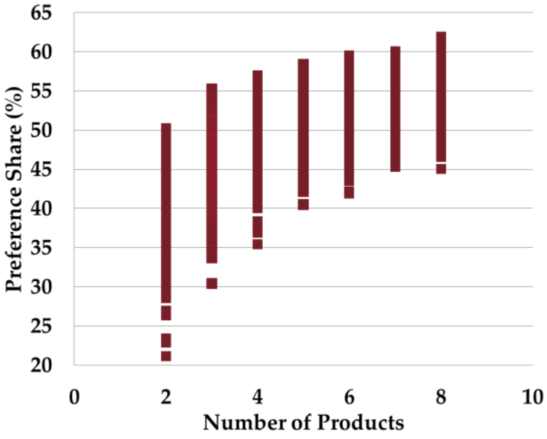


Figure 13. Share of preference distribution for the targeted initial population

The results of Figure 13, like those of Figure 6, show that as product line size increases, the preference share distribution tightens. However, unlike the first case study problem, there is no apparent limit to the preference share that can be captured. As more products are added to the product line, the best solution captures more share than the best solutions of smaller product lines.

The number of objective function evaluations required to construct the targeted population (corrected to market level using Equation 4) is shown in Table 13. For small product lines, sub-optimization with subsets of respondents provides adequately sized pools of optimal vehicles for generating targeted populations of initial product lines. Hence the number of required objective function calls steadily increases as the product line size increases from 2 to 5 products. Once the product line size reaches 5 products, sub-optimization with the entire population of respondents is required to generate diverse targeted populations. At this point, no further computational costs are incurred in generating the targeted populations; however, some optimal products must be reused to fill the initial population strings.

Table 13. Objective function evaluations needed to create the targeted initial population

Number of products	2	3	4	5	6	7	8
Number of evaluations	399	897	1588	2236	2236	2236	2236

In the first case study, it was shown that the targeted population initially outperformed the randomly generated population. However, when the random population was allowed to undergo genetic operations to match the computational cost of constructing the targeted population, it was observed that the advantage of the targeted population was limited to larger product line sizes. However, for this case study problem, the design space is significantly larger. Performing the same analysis as in the first case study problem, the best solutions from 5 runs of the targeted and randomly generated initial populations are shown in Figure 14. For the second case study, the targeted population demonstrates a significant performance advantage across all product line sizes when an equivalent number of function evaluations are considered.

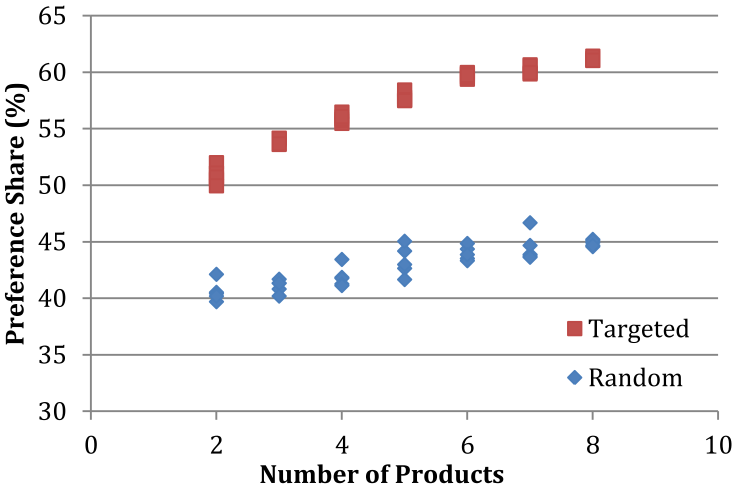


Figure 14. Performance comparison at equivalent objective function evaluations

Figure 15 compares the best solution of the targeted initial population at creation to the best solution for the Sawtooth Software GA when the stopping criterion has been satisfied. This result is significant in that the best solution from the targeted population outperforms the best solution from the Sawtooth Software GA (run to completion) at every product line size. When the targeted population is allowed to run to completion, the quality of the solution further increases. This result is shown in Figure 16.

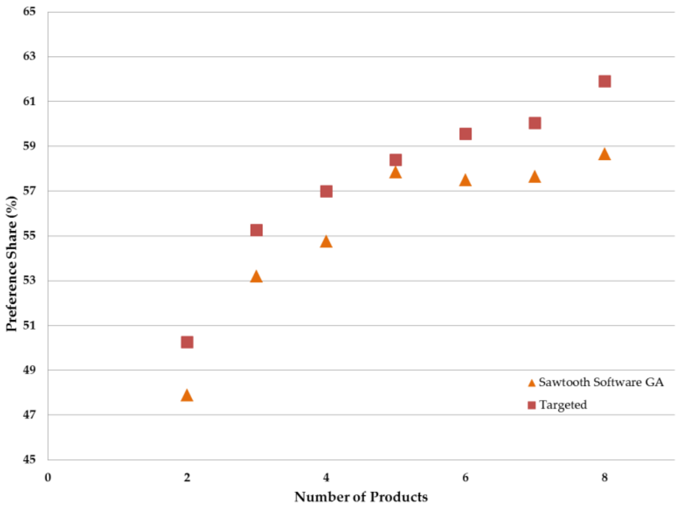


Figure 15. First comparison of targeted population and Sawtooth Software GA

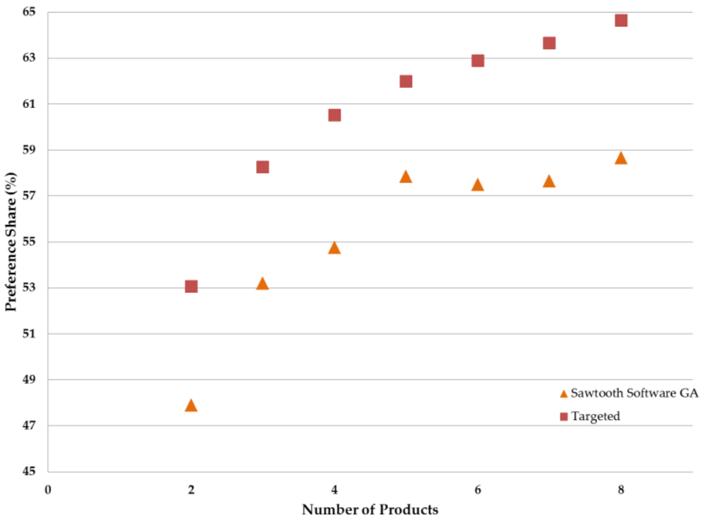


Figure 16. Comparison of targeted population and Sawtooth Software GA when both are run to completion

As in the first case study, the results in Figure 16 demonstrate that the best product line solution achieved by the targeted population outperforms the best product line solution found by the Sawtooth Software GA. Additionally, preference share increases monotonically as the size of the product line increases. This trend is not exhibited in the solution found by the Sawtooth Software GA.

To compare the computational costs of each approach, the numbers of evaluations required by the Sawtooth Software GA are shown in Table 14. Table 15 shows the numbers of function calls required by the targeted population approach to match the best solution obtained by the Sawtooth Software GA along with the numbers of evaluations required to satisfy the stopping criterion. Since the targeted population outperforms the Sawtooth Software GA upon creation, the number of function evaluations required to match performance equals the computational cost of creation.

Table 14. Objective function evaluations needed for the Sawtooth Software GA to satisfy stopping criterion

Number of products	2	3	4	5	6	7	8
Number of evaluations	7000	10800	16400	22200	21000	22500	22500

Table 15. Objective function evaluations needed for the targeted initial population genetic search in a Matlab environment

Number of products	2	3	4	5	6	7	8
Evaluations to match	399	897	1588	2236	2236	2236	2236
Total evaluations	6800	13500	28800	52236	59856	129636	83036

To further explore the computational savings offered by the targeted population, solutions from the Sawtooth Software GA are again compared to results from genetic searches run using random initial populations within the Matlab environment. Figure 17 shows the best performance obtained by both methods when allowing the same number of objective function evaluations. Under these conditions, it is observed that a random population genetic search in Matlab performs better for smaller product lines while the Sawtooth Software GA performs better for larger product lines.

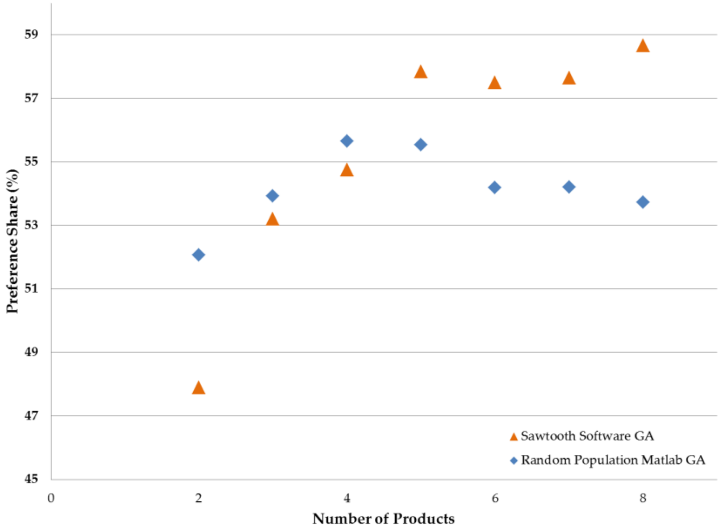


Figure 17. Comparison of random population GA and Sawtooth Software GA

However, when both methods are run to completion, the random population GA returns solutions similar to those found using the targeted population GA. This is shown in Figure 18.

Interestingly, this trend does not hold for smaller product line solutions; this phenomenon will be explored in future work. For larger populations, the convergence of both methods to similar solutions shows that the random population GA is also capable of outperforming the Sawtooth Software GA (i.e. the improvement in solution quality shown in Figure 15 applies to both the random and targeted initial population GAs). The targeted population may provide marginal benefits in solution quality, but as shown in Table 16, the primary benefit of the targeted population is reduction of computational cost.

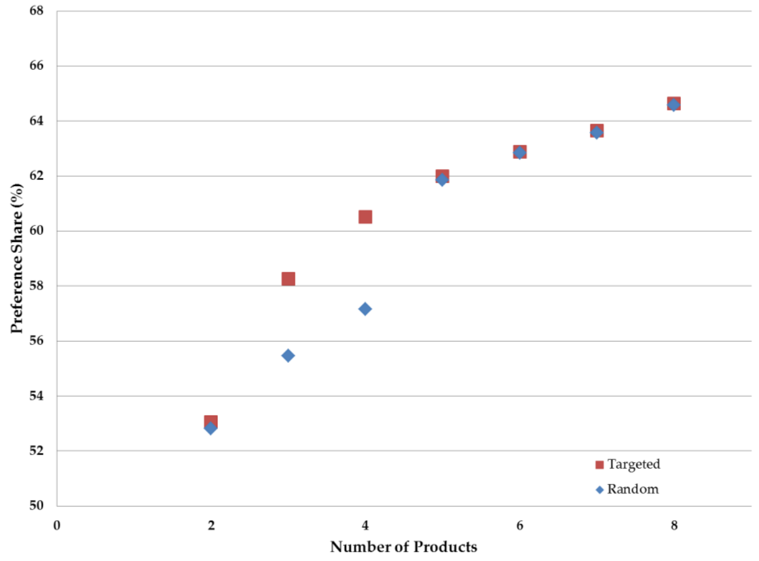


Figure 18. Comparison of random population GA and targeted population GA at completion

Table 16. Total number of objective function evaluations needed for the random population GA and the targeted population GA

Number of products	2	3	4	5	6	7	8
Random population	12400	23400	40800	149000	100200	141400	162400
Targeted population	6800	13500	28800	52236	59856	129636	83036

4.3 Discussion of results

Two case study problems were presented in this section to demonstrate the scalability of the targeted initial population approach. While application of a targeted population yielded quantifiable benefits for a small-scale problem, the true power of this approach was seen in a problem with a large solution space. In both problems, the targeted population GA returned solutions with preference shares equivalent to the Sawtooth Software GA solutions in fewer objective function evaluations. Also, when allowed to run until convergence criteria were satisfied, the targeted population solutions consistently outperformed the Sawtooth Software GA solutions; the average improvements in preference share were 1 percent improvement in the MP3 player problem and more than 5 percent in the automobile feature packaging problem. Results obtained from the targeted initial population approach are also directionally consistent, exhibiting the expected trend of increasing preference share as more products are introduced, in contrast to the Sawtooth Software GA solutions shown in Figure 1. This suggests that techniques capable of reducing the computational complexity of a problem lead to better solution quality.

In this section it has been demonstrated that the targeted initial population approach may be applied to generate higher-quality solutions with lower computational costs. In addition, genetic searches offer an additional advantage of extensibility to problems involving multiple objectives. The following section demonstrates how the targeted initial population approach may be effectively adapted for application to multiobjective optimization problems.

5. Extension to multiobjective optimization problems

Companies may not wish to determine their market positioning strategy based only on share of preference. Often preference share may be easily increased by reducing prices across the entire product line; however, this action typically has a negative impact on overall profitability. Thus, there is an inherent tradeoff between preference share and profit that must be explored.

In many engineering optimization problems, there are multiple criteria that a product designer wishes to maximize or minimize. When each criterion is represented as an objective function, the multiobjective problem formulation shown in Equation 5 may be applied:

$$\begin{aligned} \min\ & f_1(x), f_2(x),..., f_n(x) \\ s.t.\ & g(x) \leq 0 \\ & h(x) = 0 \\ & x_{lb} \leq x \leq x_{ub} \end{aligned} \tag{5}$$

For packaging problems, *x* represents the vector of product attributes comprising a product line, *g(x)* are general inequality constraints, and *h(x)* are general equality constraints. In this formulation, there are *n* competing objective functions, f_1 through f_n, that need to be minimized.

In problems with multiple competing objectives, the optimum is no longer a single solution but an entire set of non-dominated solutions. This set of non-dominated solutions is commonly known as the “Pareto set” and is comprised of Pareto dominant solutions [34]. A solution is said to be non-dominated if there are no other solutions that perform better on at least one objective and perform at least as well on the other objectives. A product attribute vector $\bar{x}$ is said to be Pareto optimal if and only if there does not exist another vector $\overline{x'}$ for which Equations 6 and 7 hold true. In other words, a solution is considered to be Pareto optimal if no objective can be improved without changing at least one other objective.

$$f_i(\bar{x}') \leq f_i(\bar{x}) \ \ \textit{for i = 1 .. n} \tag{6}$$
$$f_i(\bar{x}') < f_i(\bar{x}) \ \ \textit{for at least one i, } 1 \leq i \leq n \tag{7}$$

Sawtooth Software’s ASM does not currently support the simultaneous optimization of multiple objectives. However, multiobjective genetic algorithms (MOGAs) are a natural extension of the traditional genetic search and are frequently applied in the optimization of complex systems. Examples of MOGA applications in product line design include the simultaneous optimization of weight and efficiency and balancing the tradeoff between commonality and distinctiveness. Advantages of the MOGA include reducing the number of function evaluations needed to converge to a set of solutions (relative to other methods such as grid searches or iterative weighted sums) and robustness to ill-conditioned problems (such as those with discrete variables or those with multimodal or discontinuous responses).

The next section of this paper describes the steps that are necessary to extend the targeted initial population approach to multiobjective product line design. The automobile feature packaging problem introduced in Section 4.2 will be used to illustrate the challenges and outcomes of this extension.

5.1 Formulating the multiobjective problem and finding a baseline solution

For the purposes of this demonstration, two objective functions are simultaneously considered. The first objective function is maximization of the preference share captured by the product line; this is identical to the objective function used throughout this paper. The second objective function is based on a profitability metric known as contribution margin. This is simply the difference between the product's selling price and its cost; it does not account for investments, other fixed costs, or the time value of money. When this measure is aggregated across all products in a product line, it is known as aggregate contribution margin (ACM). When expressed as an objective function, it is (obviously) to be maximized.

Beyond addition of the second objective, the packaging problem formulation is further extended to include pricing structure as a decision variable. This extension would be meaningless in the previous problem formulation; when preference share is the only objective, it may be safely assumed that lowering price will have a favorable effect. The addition of ACM as an objective function introduces tension between the customer-based objective of maximizing probability of choice and the company-based objective of maximizing profit. Pricing structures have been established for each attribute level based on historical costs and market price thresholds determined by subject matter experts at General Motors. A representative pricing structure for a hypothetical product attribute is shown in Table 17.

Table 17. Representative price structure for a hypothetical product attribute

Attribute level	Cost	Low price bound	High price bound
1	$0	$0	$20
2	$50	$60	$120
3	$125	$125	$250

Final extensions of the problem formulation are the identification of standard / optional equipment and the addition of trim levels (e.g. LS, LT, LTZ). Multiple vehicles may exist in each trim level. The first product in each trim level represents the standard equipment offered for that trim level; any content changes for the other vehicles in the trim levels are considered to be additions of optional equipment. Standard equipment and optional equipment are allowed to have different price structures.

The trim levels also function as anchor points on the price scale around which different sets of products may be offered. As shown in Table 18, five different trim levels are considered in this example, containing a total of 19 different products that must be configured. The full multiobjective problem formulation is shown in Equation 8.

Table 18. Number of products per trim level

Trim level	1	2	3	4	5
Products per trim	3	3	4	5	4

Maximize: *Preference share*

ACM

with respect to: *Feature content*

Feature prices (standard and optional)

Standard and optimization equipment (8)

subject to: *Feature price bound*

Number of trim levels

Number of vehicles per trim

Trim level price bounds

The optimization problem formulated in Equation 8 contains a total of 481 design variables – 361 to represent 19 attribute levels on each of 19 products plus an additional 120 to cover different markup levels for standard and optional equipment. For the objective functions, preference shares are reported as raw numbers while values of ACM are normalized with respect to a pre-defined baseline configuration for reporting.

To establish a baseline solution for this problem formation, a multiobjective genetic algorithm (MOGA) was executed with a random initial population using the input parameters shown in Table 19. This MOGA was an extension of the genetic algorithm coded in MATLAB that was discussed in Section 4. Figure 19 depicts the location of the frontier when the stopping criterion was achieved; this required 700 generations and 660,000 objective function evaluations.

Table 19. Input parameters the MOGA

Criteria	Setting
Pool size	1000
Offspring created within a generation	(1/2)*(pool size)
Selection	Random
Crossover type	Uniform
Crossover rate	0.5
Mutation type	Gaussian
Mutation rate	0.5
Stop after	User defined number of generations

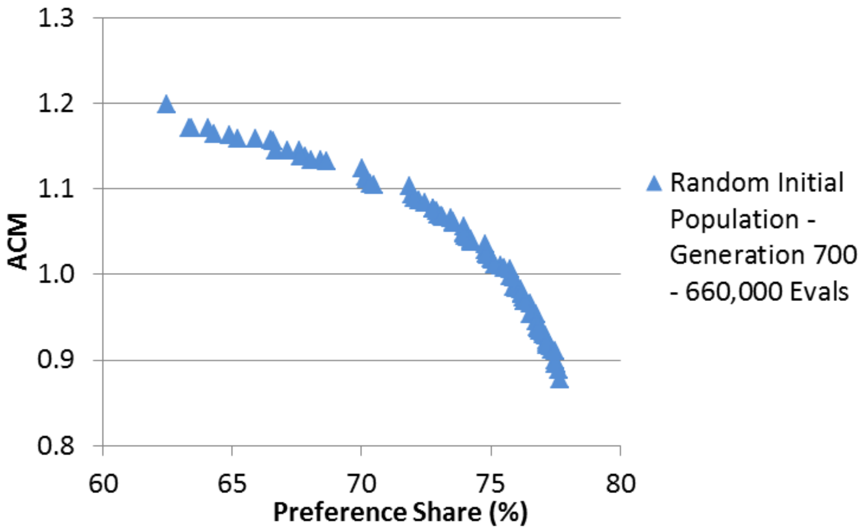


Figure 19. Multiobjective solution of the random population

Developing a suitable method for managing the interactions between the price structure and the trim level strategy posed an additional challenge in creating a targeted initial population. As described in Section 3, development of the targeted population involved identification of an optimal product configuration for each respondent. The sub-optimization procedure is straightforward with a fixed structure; however, when feature prices are allowed to vary, solutions for optimal product configurations will also vary. Failure to address this issue when generating the targeted initial population has serious ramifications for solution quality. Figure 20 shows the frontier obtained when the targeted initial population is generated by fixing the prices at the midpoints of their respective scales. The algorithm converges to a frontier similar to that obtained from the random initial population but saved 100 generations and 95,000 objective function evaluations.

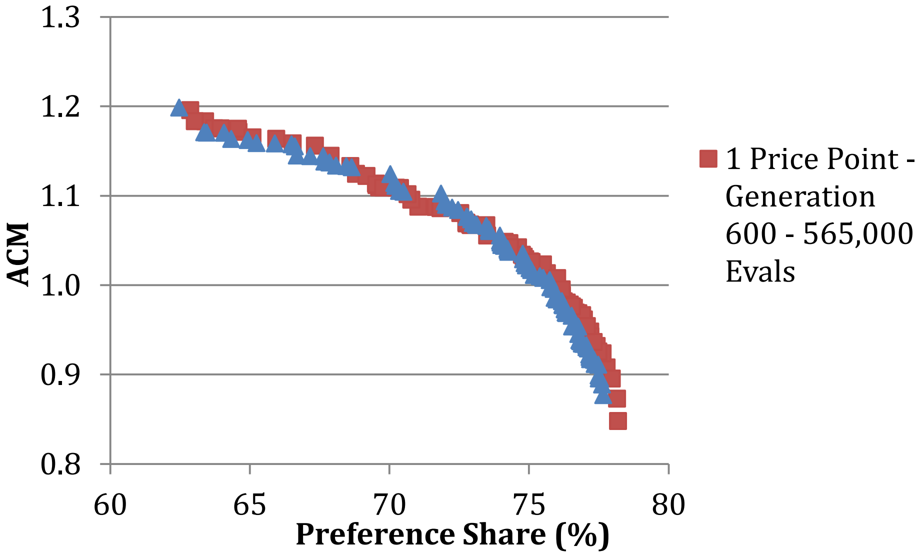


Figure 20. Multiobjective targeted frontier using only a mid-range price point

As shown in Figure 21, fixing feature prices at the midpoints of their respective price scales severely limits the diversity of the non-dominated targeted population seed solutions in the multiobjective performance space. Although the targeted initial population converged to the same performance space in 20% fewer function calls, additional benefits can be realized by achieving a greater diversity of seed solutions for the targeted initial population. To achieve this greater seed diversity in the multiobjective performance space, it is desirable to adapt the method by which the targeted initial populations are generated. This is described in the next section.

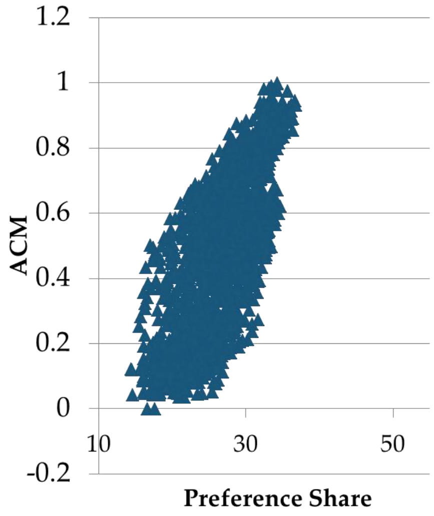


Figure 21. Multiobjective representation of targeted initial population using only a mid-range price point

5.2 Adapting the targeted initial population generation process

To increase the diversity of the initial population in the multiobjective performance space, an additional layer of sub-optimization has been added to the initial population generation method. First, the price scales for each attribute level were discretized with equal numbers of increments; as shown in Figure 22, for this case study the price scales were divided into quartiles to generate a total of 5 price points. Next, a sub-optimization is performed to identify an optimal vehicle for each respondent at each price point within each trim level. With 5 price points and 5 trim levels used in this case study, 25 targeted product solutions were generated for each respondent.

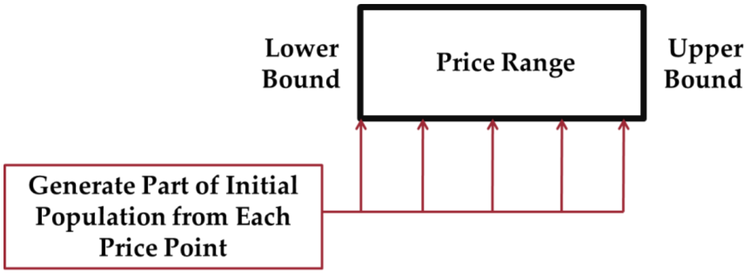


Figure 22. Adaption to the development of targeted population solutions

The effectiveness of this procedure is shown in Figure 23. The diversity of solutions in the initial population has increased dramatically. More importantly, the number of non-dominated solutions in the targeted population points has increased substantially. These increases in diversity and non-dominance would be expected to translate to higher quality final solutions and further reductions in computational expense. The potential for these outcomes is explored in the next section.

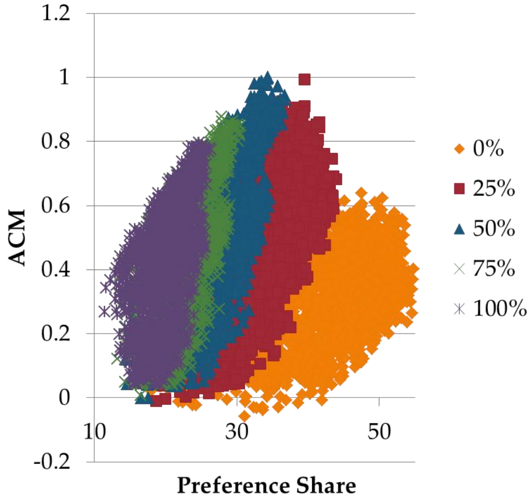


Figure 23. Multiobjective representation of targeted initial population using five price points

5.3 Comparing multiobjective optimization solutions

The two multiobjective solutions are shown in Figure 24. Here, it is seen that the modified targeted initial population solution completely dominates the solution obtained by the randomly generated population. Further, the modified targeted initial population is able to achieve this result using 121,000 fewer objective function evaluations.

Having demonstrated the effectiveness of targeted initial populations for both single and multiobjective problems, the next section revisits the conclusions and poses areas of future work.

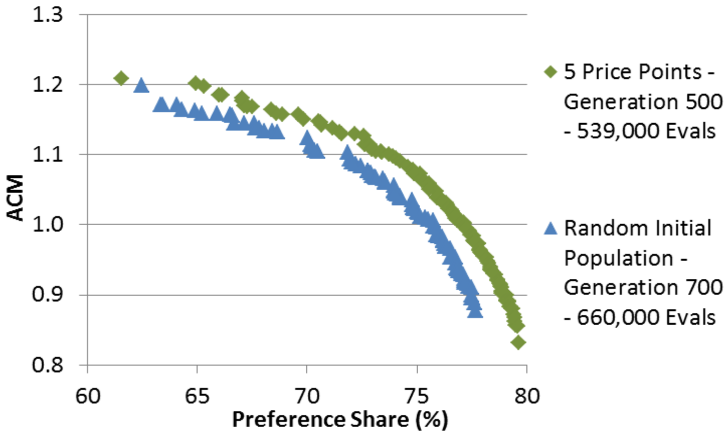


Figure 24. Comparison of non-dominated solutions

6. Conclusions and future work

The goal of this investigation was to explore the benefits of applying new genetic algorithm technologies to product search methods with discrete choice models. As anticipated, application of a targeted initial population yielded a number of benefits, including reductions in the computational cost required to complete the product searches, improvements in the preference shares for the solutions from the product searches, and more consistent performance of product search solutions on problem objectives. These benefits scaled favorably as problem size increased; preference share increased monotonically as product variants were added and improvements in genetic algorithm performance due to the targeted initial population increased geometrically with increased problem size. Additionally, this work successfully incorporated product feature pricing structure as a variable in the product search; in this case, the performance of the product search algorithm was substantially improved by varying feature price structures during generation of the targeted initial population.

Implementation of a multiobjective genetic algorithm provided the ability to explore trade-offs between competing business objectives; the trade-off between preference share and aggregate contribution margin was explored in this work although other objectives could also be considered.

Future work will include extension of the core genetic algorithm technology as well as generalization of its application. Alternative sub-optimization procedures will be explored for generating the targeted initial population in hope of improving computational efficiency. The repeatability of genetic algorithm solutions will be examined at the product configuration level. The effects of varying genetic algorithm parameters such as the mutation and crossover rates and initial population size on performance metrics including achievement of objectives, runtime, and solution diversity will be explored in detail. Development of a run-time advisor capable of estimating minimum effective runtimes, runtime to exhaustion, and potential payoffs of continued analysis (measured in terms of objective function improvement) will be investigated.

Finally, extension of this approach will be pursued by seeking collaborative research opportunities in three primary areas: integration with synergistic research efforts in product platforming and system architecting, application to a more diverse portfolio of conjoint product searches, and generalization of the methods for application to additional classes of optimization problems.

ACKNOWLEDGEMENTS

We gratefully acknowledge support from the National Science Foundation through NSF Grant No. CMMI-0969961. Any opinions, findings, and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Arora, J, S., 2004, Introduction to Optimum Design – 2nd edition, Academic Press.

[2] Rao, S., 2009, Engineering Optimization: Theory and Practice – 4th edition, Wiley.

[3] Sawtooth Software, 2003, “Advanced Simulation Module for Product Optimization v1.5 Technical Paper, Sequim, WA.

[4] Holland, J., 1975, “Adaptation in Natural and Artificial Systems,” *SIAM Review,* **18**(3): 287-299.

[5] Michalek, J., Feinberg, F., and Papalambros, P., 2005, “Linking Marketing and Engineering Product Design Decisions via Analytical Target Cascading,” *Journal of Product Innovation Management,***21**(1): 42-62.

[6] Camm, J., Curry, D., Cochran, J., and Kannan, S., 2006, "Conjoint Optimization: An Exact Algorithm for the Share-of-Choice Problem," *Management Science,* **52**(3): 435-447.

[7] Balakrishnan, P. V., Gupta, R., and Jacob, V. S., 1996, “Genetic Algorithms for Product Design,” *Management Science*, **42**(8): 1105-1117.

[8] Besharati, B., Luo, L., Azarm, S., and Kannan, P. K., 2004, “An Integrated Robust Design and Marketing Approach for Product Design Selection Process,” *2004 ASME IDETC Conference,* Salt Lake City, UT, DETC04/DAC57405.

[9] Besharati, B., Luo, L., Azarm, S., and Kannan, P. K., 2006, “Multi-Objective Single Product Robust Optimization: An Integrated Design and Marketing Approach,” *Journal of Mechanical Design,* **128**(4): 884-892.

[10] Green, P. E., and Krieger, A. M., 1985, “Models and Heuristics for product Line Selection,” *Marketing Science,* **4**(1): 1-19.

[11] Sudharshan, D., May, J. H., and Shocker, A. D., 1987, “A Simulation Comparison of Methods for New Product Location,” *Marketing Science*, **6**(Spring): 182-201.

[12] Kohli, R., and Krishnamurti, R., 1987, “A Heuristic Approach to Product Design,” *Management Science*, **33**(12): 1523-33.

[13] Dobson, G., and Kalish, S., 1993, “Heuristics for Pricing and Positioning a Product-Line using Conjoint and Cost Data,” *Management Science,* **39**: 160-175.

[14] Nair, S. K., Thakur, L. S., and Wen, K. W., 1995, “Near Optimal Solutions for Product Line Design and Selection: Beam Search Heuristics,” *Management Science,* **41**: 767-85.

[15] Thakur, L. S., Nair, S. K., Wen, K. W., and Tarasewich, P., 2000, “A New Model and Solution Method for Product Line Design with Pricing,” *Journal of the Operational Research Society,* **51**: 90-101.

[16] McBride, R. D., and Zufryden, F., S., 1988, “An Integer Programming Approach to the Optimal Product Line Selection Problem,” *Marketing Science,* **7**(2): 126-140.

[17] Hanson, W., and Martin, K., 1996, “Optimizing Multinomial Logit Profit Functions,” *Management Science,* **42**(7): 992-1003.

[18] Michalek, J., Ceryan, O., Papalambros, P., Koren, Y., 2006, “Balancing Marketing and Manufacturing Objectives in Product Line Design,” *Journal of Mechanical Design,* **128**(6): 1196-1204.

[19] Wang, X., Camm, J., and Curry, D., 2009, "A Branch-and-Price Approach to the Share-of-Choice Product Line Design Problem," *Management Science,* **55**(10): 1718-1728.

[20] Li, H., and Azarm, S., 2002, “An Approach for Product Line Design Selection Under Uncertainty and Competition,” *Journal of Mechanical Design,* **124**(3): 385-392.

[21] Chapman, C., and Alford, J., 2011, “Product Portfolio Evaluation Using Choice Modeling and Genetic Algorithms,” *Proceeding of the 2010 Sawtooth Software Conference,* Newport Beach, CA.

[22] Steiner, W., and Hruschka, H., 2003, “Genetic Algorithms for Product Design: How Well Do They Really Work?” *International Journal of Market Research,* **45**(2): 229-240.

[23] Belloni, A., Freund, R. M., Selove, M., and Simester, D., 2008, “Optimal Product Line Design: Efficient Methods and Comparisons,” *Marketing Science,* **54**(9): 1544-1552.

[24] Tsafarakis, S., Marinakis, Y., and Matsatsinis, N., 2011, “Particle Swarm Optimization for Optimal Product Line Design,” *International Journal of Research in Marketing,* **28**: 13-22.

[25] Balakrishnan, P. V., Gupta, R., and Jacob, V. S., 2006, “An Investigation of Mating and Population Maintenance Strategies in Hybrid Genetic Heuristics for Product Line Designs,” *Computer & Operations Research,* **33**(3): 639-659.

[26] Spears, W., and DeJong, K., 1991, “An Analysis of Multi-Point Crossover,” *Foundations of Genetic Algorithms,* G. Rawlins, ed. Morgan-Kaufmann.

[27] Correa, E., Steiner, M., Freitas, A., and Camieri, C., 2001, “A Genetic Algorithm for the P-Median Problem,” *Genetic and Evolutionary Computation Conference,* GECCO-2001, San Francisco, CA.

[28] Poles, S., Fu, Y., and Rigoni, E., 2009, “The Effect of Initial Population Sampling on the Convergence of Multi-Objective Genetic Algorithms,” *Multiobjective Programming and Goal Programming,* **618**: 123-133.

[29] Sawtooth Software, 2008, “CBC v.6.0”, Sawtooth Software, Inc., Sequim, WA, http://www.sawtoothsoftware.com/download/techpap/cbctech.pdf.

[30] Sawtooth Software, 2009, “The CBC/HB System for Hierarchical Bayes Estimation Version 5.0 Technical Paper,” Sawtooth Software, Inc., Sequim, WA, http://www.sawtoothsoftware.com/download/techpap/hbtech.pdf.

[31] Goldberg, D. E., 1989, Genetic Algorithms in Search, Optimization & Machine Learning, Addison-Wesley.

[32] Turner, C., Ferguson, S., and Donndelinger, J., 2011, “Exploring Heterogeneity of Customer Preference to Balance Commonality and Market Coverage," ASME Design Engineering Technical Conference, Design Automation Conference, Washington, DC., DETC2011-48581.

[33] Matlab, the Mathworks.

[34] Pareto, V., 1906, Manuale di Econòmica Polìttica, Società Editrice Libràia, Milan, Italy; translated into English by A. S. Schwier, as Manual of Political Economy, Macmillan, New York, 1971.

Discussion of Ferguson, Turner, Foster, Donndelinger, and Beltramo

Christopher N. Chapman
GOOGLE

Overview

The paper by Ferguson, Turner, Donndelinger, and Beltramo is significant for two reasons: the authors present an important application of genetic algorithms (GAs) for product line optimization in the automotive industry and they show how a common concern with GAs might be addressed by using information from individual-level utility estimates from conjoint analysis.

Today's computing resources make it possible to perform search and optimization processes that would have required prohibitive amounts of time just a few years ago. The search space of all combinations in a product line can be many orders of magnitude too large to be searched by exhaustive or simple procedures. Complex search spaces demand heuristic algorithms; GAs have been shown to perform well, often approaching optimal results, across many kinds of heuristic search problems (Goldberg, 1989; Mitchell, 1998).

In marketing, Belloni *et al.* (2008) demonstrated that GAs could perform well at finding product lines from conjoint analysis data. In 2009, my team (Chapman & Alford, 2010) used GAs to find the optimal size of a consumer product line and to determine which products should be cut or retained from a large line in actual production. This led to substantial simplification of the line with attendant cost savings. The analysis also revealed products that were highly desirable to consumers but that no manufacturer was making.

In the present paper, Ferguson *et al.* considered a general concern with GAs: a GA exhibits some degree of dependence on the initial population. In the case of product line optimization, if a product comprises randomly selected attributes, it is likely to have very low fitness. Thus, a GA that starts with a population of such products will take longer to find a high-performing solution. The authors propose a solution: seed the initial population with products that are optimal at an individual level by selecting features on the basis of individual-level utilities from a conjoint study. They call this a “targeted initial population” (TIP), and show in the automobile case that a GA with population initialized in this manner achieves high performance faster than a GA without TIP.

This is an interesting and useful suggestion for GAs, and is an intriguing way to use conjoint results to inform the GA rather than treating GA as a generic optimizer. However, before recommending TIP as a general procedure for practitioners, I wish to review a few aspects of GA options that affect optimization and run time. I intend this primarily to provide discussion points and guidelines for practitioners.

The Many Levers of GAs

- **Population size**: there is often an inverted U curve relationship between population size and performance; increasing population size leads to better performance up to a point, and after that incurs an evaluation cost for little improvement. A general rule of thumb for static, vector-based genome structures is to use a population of 200-400 candidates.

- **Mutation rate**: again, this is U shaped: mutation is necessary or else an initial population is simply recombined with no change; yet too much mutation will wipe out the gains of previous generations. A rule of thumb is a mutation rate of 0.005-0.01 per gene for typical fixed-length genomes (20-100 integer or real number genes), and to consider a declining mutation rate across generations.

- **Fitness function partial credit**: GAs perform better if the fitness function is not *all-or-nothing* but allows partial credit for solutions that mix good and bad elements. Luckily, this is almost automatically the case in many marketing applications with conjoint, which score population interest. Check: if your fitness is zero (or some other fixed value, or hits a ceiling) for many candidates, rework it with a partial credit or stochastic scoring scheme.

- **Crossover rate and method**: the process of recombining genomes is essential to a GA and there are many available methods (single point, multi-point, all points, inversion, etc.). A default approach is to have a high crossover rate, always preserve the best individual candidate (elitism), and use several crossover methods concurrently instead of a single method.

- **Run time**: the longer a GA runs, the more likely it is to find an optimal outcome. Some GA libraries now offer multicore and multimachine options, which allow a linear increase in *effective* run time with little cost in real time. (See references.)

- Finally, one might consider **more elaborate procedures**, such as mapping a higher-order genome structure onto a translation that is the phenotype; or having a non-fixed genome structure, where the ultimate case is genetic programming to evolve a variable-length program.

In short, if a GA requires substantial customization and a lengthy run time – as most do, because otherwise one might as well use an exhaustive search – then one should experiment with GA parameters and not assume that a default implementation will be particularly useful.

These considerations show one limitation of the authors' paper: they use the Sawtooth Software SMRT genetic algorithm product line search, which does not afford as many options as a custom GA. Rich Johnson pointed out in live Q&A at the conference that the authors' comparison against the SMRT GA is difficult to assess. Their model optimized faster than SMRT, but it is unclear whether minor tweaks to the Sawtooth Software model, such as running longer, could have similar performance.

Discussion of Targeted Initial Populations

How do the GA choices above relate to targeted populations (TIPs)? First, if ***mutation rate*** is high enough (but not too high) and ***run times*** are long enough, then in principle any population is equally likely to converge on an optimum. This suggests that TIPs would be of primary value in cases where run time is a concern, as may be the case for large-scale industrial applications. An alternative to TIP would be to increase effective run times with multiprocessor/multimachine GA libraries.

Second, TIPs should work best where there is substantial (real, i.e., human) ***population heterogeneity*** in per-individual optima, i.e., substantially different individual preferences. A GA needs diversity in genes, or else it is wasting evaluation cycles on too-similar candidates. With conjoint data, if most or all respondents agree on an attribute value (e.g., price), then a naively implemented TIP might have little diversity. In such a case, one could select a TIP using a data source with more heterogeneity such as hierarchical Bayes (HB) draws of individual estimates.

Third, if the fitness landscape is complex, with prohibitions or linked attributes, then ***finding an individual optimal product*** may be a non-trivial search process in itself. In other words, finding a TIP would be a difficulty, not a shortcut. In such a case, one might decide to perform a partial instead of optimal TIP search; or skip TIP altogether and compensate with a larger run time or population.

Fourth, TIP candidates could easily be ***combined*** with random candidates in an initial population, and there is no general reason not to do so. If one is able to find TIPs conveniently, a conservative strategy for population construction might be to generate proportion *p* (e.g., 0.5) of the population at random, and another proportion *1-p* using TIP. This is especially appealing when a TIP-like optimum is readily available, such as the data from a *build-your-own* conjoint task.

In conclusion, the work here is valuable for both the case presentation and the authors' suggestion of a new method to improve the odds that a GA starts well and is likely to reach a near-optimal solution. As computing power advances, and heuristic algorithms are able to solve larger and larger problems, I am confident that the future of marketing research will have many applications for GAs. The combination of GA search with conjoint analysis data is an extremely appealing method for marketing practitioners because it allows us to examine difficult but very interesting strategic problems.

Annotated References

1. A. Belloni, R. Freund, M. Selove, & D. Simester (2008). Optimizing product line designs: efficient methods and comparisons. *Management Science 54:9, 1544-1552*.
Shows that GA model can come very close to optimal results in portfolio optimization problems using conjoint analysis data, with less complexity or risk than some other algorithms.

2. C.N. Chapman (2010). Genetic algorithms for choice model portfolio analysis. [R code]
Computer software; R code to find near-optimal product lines using GA with utilities from CBC, ACBC, or similar conjoint data, with options for using HB draws. Available by emailing this author.

3. C.N. Chapman & J.L. Alford (2010). Product portfolio evaluation using choice modeling and genetic algorithms. *Proc. Sawtooth Software Conference 2010.*, Newport Beach, CA.
Case example with GA + ACBC to inform major strategic reorganization of a real product line.

4. D.E. Goldberg (1989). *Genetic algorithms in search, optimization, and machine learning*.
The standard introduction to GAs, a leisurely read and not up to date, yet still a classic.

5. W.R. Mebane & J.S. Sekhon (2011). rgenoud. http://sekhon.berkeley.edu/rgenoud/
A *GA library for R, implemented in Fortran for speed, with multiprocessor & multimachine options.*

6. M. Mitchell (1998). An introduction to genetic algorithms. Cambridge, MA: MIT Press.
A concise introduction to GAs with interesting examples and guidance on parameters.

7. R. Poli, W.B. Langdon, & N.F. McPhee (2008). A field guide to genetic programming. Lulu.com.
A practical introduction to genetic programming models, which extend GA to abstract and flexible models

8. M. Wall (1999). GAlib. http://lancet.mit.edu/ga/
A *well-designed, single-threaded GA library for C++; dated but still useful.*

CAN WE IMPROVE CBC QUESTIONNAIRES WITH STRATEGICALLY-PLACED LEVEL OVERLAP AND APPROPRIATE "SCREENING RULE" QUESTIONS?

KEVIN LATTERY
MARITZ RESEARCH
BRYAN ORME
SAWTOOTH SOFTWARE

1. INTRODUCTION

In 2007, Johnson and Orme presented a new method for ACBC, which has been commercialized by Sawtooth Software. It involves the three-stage process of BYO + Consideration + Choice Tournament. One of the characteristics of this approach is that after observing a pattern of choice behavior in the Consideration stage, the software probes respondents regarding whether consistently-rejected levels are indeed unacceptable. If respondents confirm cutoff rules, then subsequent choice tasks will only include acceptable levels. Some researchers have wondered whether the BYO and Consideration phases may be dropped, and CBC-only tasks be used in a similar adaptive process. We call this test version adaptive DCM.

We compare this adaptive approach with a non-adaptive approach, employing non-adaptive experimental designs and stated information about the levels of each attribute. For the non-adaptive approach, we tested two experimental designs. One of these designs was a simple D-efficient design with little overlap. The second design had more overlap.

We conducted a fairly rigorous test of the adaptive DCM and two fixed designs. This involved separate cells for each design and a holdout cell. We also included tasks used for calibrating the scale parameter and holdout hit-rate. The next section discusses the details of this test.

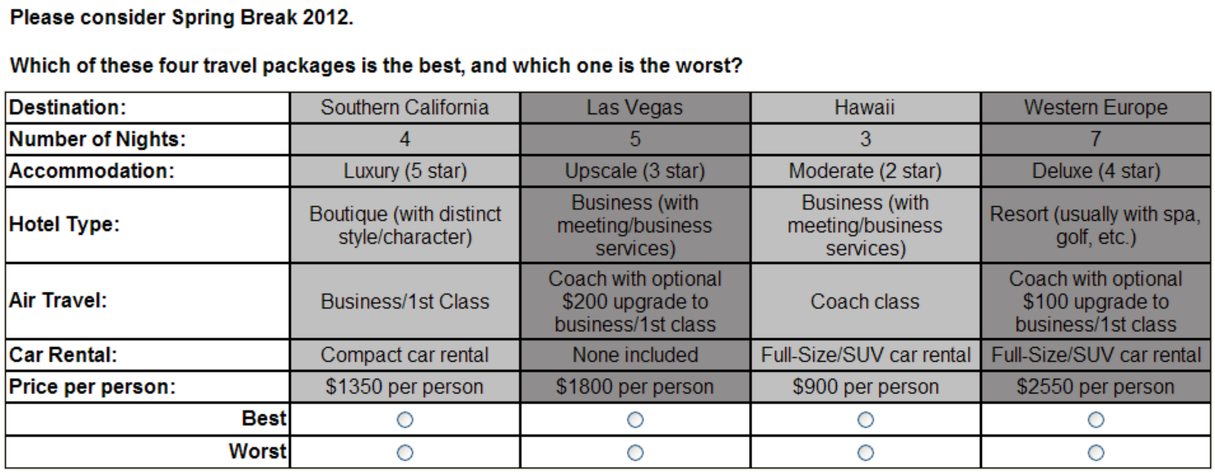
2. DETAILS OF RESEARCH DESIGN

Respondents were randomly split, in real time, into one of four cells. For every four respondents that clicked on a link to participate, one was sent to Sawtooth Software's web site to do an adaptive DCM. The other three were sent to a Critical Mix website to do one of three fixed designs: low overlap, high overlap, or holdout.

A) Three Test Cells

All three test cells (adaptive DCM, low overlap, high overlap) showed the respondent 12 tasks. The low overlap and high overlap fixed designs had a total of 60 tasks, with 5 blocks of 12 tasks. Respondents were assigned to do one block based on whichever block had the fewest completed records.

All test cells appeared very similar to one another in formatting, and were identical in attribute and level wording. The tasks showed 4 alternatives and looked like this:



Please consider Spring Break 2012.

Which of these four travel packages is the best, and which one is the worst?

Destination:	Southern California	Las Vegas	Hawaii	Western Europe
Number of Nights:	4	5	3	7
Accommodation:	Luxury (5 star)	Upscale (3 star)	Moderate (2 star)	Deluxe (4 star)
Hotel Type:	Boutique (with distinct style/character)	Business (with meeting/business services)	Business (with meeting/business services)	Resort (usually with spa, golf, etc.)
Air Travel:	Business/1st Class	Coach with optional $200 upgrade to business/1st class	Coach class	Coach with optional $100 upgrade to business/1st class
Car Rental:	Compact car rental	None included	Full-Size/SUV car rental	Full-Size/SUV car rental
Price per person:	$1350 per person	$1800 per person	$900 per person	$2550 per person
Best	○	○	○	○
Worst	○	○	○	○

The low overlap and high overlap designs were constructed by attempting to maximize the D-efficiency criterion, using a modified version of Kuhfeld's SAS macros. A large set of candidate tasks were created. These were divided into low overlap, high overlap, and too much overlap. The latter group consisted of alternatives where an attribute had three or more levels that were the same. These were discarded as having too much overlap. So for the high overlap cell, the most overlap for any attribute on a specific task was two levels each appearing twice.

The following table gives the list of attributes and levels, as well as the degree of overlap for each level. There are many ways to measure overlap. In this case, we simply took the average of the number of times each level appears in a task. So no overlap gives an average of 1. For attributes with 3 levels there must be some overlap given 4 alternatives. So the minimum overlap is (1+1+2)/3 = 1.33. For the other attributes, the minimum overlap is 1.

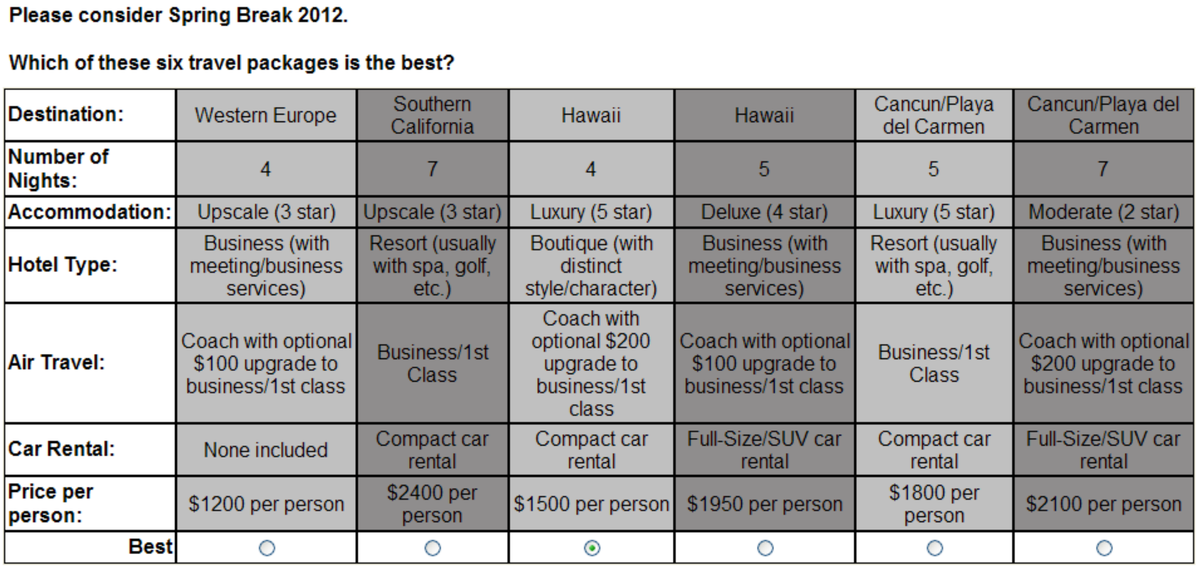
Attribute	Levels	Num Levels	Minimal Overlap	Low Overlap	High Overlap
Destination	Hawaii, Las Vegas, Orlando, Southern California, Western Europe, Cancun/ Playa del Carmen	6	1.0	1.0	1.3
Number of Nights	3 Nights, 4 Nights, 5 Nights, 7 Nights	4	1.0	1.1	1.6
Accommodation	Moderate (2 star), Upscale (3 star), Deluxe (4 star), Luxury (5 star)	4	1.0	1.1	1.7
Hotel Type	Business (with meeting/business services), Resort (usually with spa, golf, etc.), Boutique (with distinct style/character)	3	1.3	1.3	1.3
Air Travel	Coach class, Coach with optional $100 upgrade to business/1st class, Coach with optional $200 upgrade to business/1st class, Business/ 1st Class	4	1.0	1.1	1.3
Car Rental	None included, Compact car rental, Full-Size/SUV car rental	3	1.3	1.3	1.3
Price per person	$900 - $2550 Total Package Conditional on Number of Nights	5 per Night			

Preliminary research with a small convenience sample indicated that respondents tended to screen choices based on Destination, Number of Nights, and Accommodation (stars). So we set these attributes to have higher overlap in the high overlap cell. This follows Chrzan et al. (2010) who showed better results with higher overlap on attributes that respondents apply non-compensatory screening rules. Destination had 6 levels (the most), for which the design program tended to generate less overlap. So we did not get quite as much overlap in the high overlap cell for the Destination attribute. We compensated a little for this by making the low overlap cell have no overlap on Destination. There were always four different destinations in the low overlap cell. The high overlap cell typically showed 3 destinations (with one destination repeated), but occasionally showed four unique destinations (10 tasks) or only two (6 tasks). Car Rental and Hotel Type were set to have minimal overlap for both fixed design cells, since these were thought to be less important. These attributes also had some overlap even in the minimal case since they had 3 levels.

Using simulated data (with known utility values plus error), we estimated aggregate models for each of the two fixed designs. With a sample of 2000 simulated respondents, the MAE across 100 holdout tasks was 3.6% for low overlap cell and 3.5% for high overlap cell. So the designs were fairly comparable in theoretical predictive accuracy. We did not deliberately design the two cells to meet this criteria, but tested them after the design (but before fielding) just to make sure we were not generating designs of radically different efficiency.

B) Holdout Cell

The holdout cell also showed respondents 12 tasks, but showed six alternatives and asked for the best choice only. So holdout cell tasks looked like this:



Please consider Spring Break 2012.

Which of these six travel packages is the best?

Destination:	Western Europe	Southern California	Hawaii	Hawaii	Cancun/Playa del Carmen	Cancun/Playa del Carmen
Number of Nights:	4	7	4	5	5	7
Accommodation:	Upscale (3 star)	Upscale (3 star)	Luxury (5 star)	Deluxe (4 star)	Luxury (5 star)	Moderate (2 star)
Hotel Type:	Business (with meeting/business services)	Resort (usually with spa, golf, etc.)	Boutique (with distinct style/character)	Business (with meeting/business services)	Resort (usually with spa, golf, etc.)	Business (with meeting/business services)
Air Travel:	Coach with optional $100 upgrade to business/1st class	Business/1st Class	Coach with optional $200 upgrade to business/1st class	Coach with optional $100 upgrade to business/1st class	Business/1st Class	Coach with optional $200 upgrade to business/1st class
Car Rental:	None included	Compact car rental	Compact car rental	Full-Size/SUV car rental	Compact car rental	Full-Size/SUV car rental
Price per person:	$1200 per person	$2400 per person	$1500 per person	$1950 per person	$1800 per person	$2100 per person
Best	○	○	●	○	○	○

All respondents in the holdout cell saw the same 12 tasks. These tasks had a mix of overlap, from low to high.

The 12 tasks in this holdout cell were used for computing out of sample MAE and R^2. Models for each of the three test cells were estimated and used to predict tasks in the holdout cell. Those predicted shares were compared with the observed shares in the holdout cell (12 tasks x 6 alternatives = 72 shares) to compute a MAE and R^2 based on the 72 data points.

C) Calibration Tasks

MAE and R^2 vary based on scale parameter, so we tuned the scale parameter of each test cell. Since we wanted to make the three test cells as comparable as possible to one another and to predict tasks with six alternatives in the holdout cell, we tuned the scale parameter using three holdout tasks that looked like tasks in the holdout cell. These three additional tasks were asked for all respondents after their initial 12 tasks were seen. Respondents were shown a transition screen after their first 12 tasks, informing them of three additional tasks that had six alternatives, and asked for the best only.

These three tasks were used to tune the scale parameter for each test cell. They were also used for computing within-sample holdout hit-rate. Note that holdout hit-rate is not affected by any tuning of the scale parameter.

D) Sampling Methodology

We picked the travel category to get high incidence of qualified respondents. In order to qualify, respondents had to be interested in travel (Slight, Moderate or High Interest on rating scale). They also had to have traveled away from their home in the last 3 years or intended to do in the next year. This gave us respondents who were at least somewhat involved in the travel category.

To ensure that the sample was randomly split, Sawtooth Software developed a script so that for each 4 respondents who clicked a link to participate, 3 were sent to the Critical Mix site to do the fixed design survey (low overlap, high overlap, or holdout cell), and one was sent to Sawtooth Software's site to do an adaptive DCM. This ensured respondents would be randomly split, even among fast vs. slow responders.

The survey was live from Jan 6 to 10, 2012 with all 4 cells filling up concurrently. Respondent demographics were compared and were very similar to one another. Screener questions were identical. The Sawtooth Software site and the Critical Mix site had very similar formatting.

E) Summary of Design

In summary, we created 4 cells that were as closely matched as possible. There were 3 test cells, with 4 alternatives asking Best and Worst choice. There was also one holdout cell with 6 alternatives that asked best only. Finally, there were 3 holdout/calibration tasks seen by all respondents.

3. Adaptive DCM Approach

A) Description of Method

One of the methods we tested in this current research was an Adaptive Discrete Choice Measurement (A-DCM) approach[34]. For each respondent, the A-DCM used 12 CBC tasks, each with 4 concepts. The design for the A-DCM tasks initially came from a standard full-profile CBC design (generated by Sawtooth Software's Complete Enumeration approach). There were 30 questionnaire versions (blocks) within the design, where each block reflected level balance and near-orthogonality. Each task initially had minimal level overlap, meaning levels within an attribute were not repeated unless necessary to create 4 concepts per task.

Each respondent was randomly assigned to one of the 30 versions. The first 5 tasks were shown to the respondent, and within each task the respondent marked which concept was best, and which was worst. To this point, everything in the A-DCM survey has proceeded as per standard practice for full-profile, Best-Worst CBC. But, after the 5th task was answered, the interview became adaptive.

After the 5th choice task, the computer scanned the respondent's answers of "best" concepts over the previous 5 tasks, and noted whether any levels from the study design were not present in any of those 5 chosen concepts. The levels that were never present in chosen concepts were displayed to the respondent, with the following text:

[34] Although our adaptive approach has some elements that are similar to Sawtooth Software's ACBC software system, we purposefully are calling this A-DCM to avoid confusion with ACBC.

> **We've noticed that so far you've avoided trips with certain characteristics shown below. Are any of these features totally unacceptable? If so, mark the one feature that is most unacceptable, so we can just focus on trips that are a possibility for you.**
>
> ***<Show list of levels here that are potentially unacceptable. Also include an option that says "None of these are unacceptable".>***

We were worried about the tendency of respondents to be overly-aggressive in marking levels as unacceptable, so this probing screen only allowed respondents to indicate the one most unacceptable level. (Though, later opportunities would be presented for respondents to indicate more unacceptable levels.)

If the respondent indicated that a level shown in the list was unacceptable, such as "Las Vegas," then the computer looked ahead to the remaining concepts that had not yet been seen by this respondent. If "Las Vegas" was found within the concepts, then "Las Vegas" was replaced by a randomly-selected other (acceptable) level from that attribute. The original full concept including the unacceptable level was also stored for use in utility estimation (described later). Of course, all this occurred in real-time, within milliseconds, so respondents could not be aware that the content of the remaining tasks was being changed if they indicated that a level was unacceptable. The result was that if we learned that a particular level was unacceptable, we didn't bother asking respondents about that level again. And, attributes that respondents were screening on (that they were using to potentially disqualify trips from consideration) were given greater level overlap in later tasks. Thus, the likelihood of seeing multiple concepts within the task that shared the same highly preferred characteristics would increase. When this occurs, respondents need to look to secondary aspects of importance to select which concepts are best and worst.

This process of scanning the previous answers and presenting a list of potentially unacceptable levels to respondents, eliciting an unacceptable level, and replacing any unacceptable levels in the tasks not yet seen was repeated after each choice task, until respondents indicated that no more levels were unacceptable. So, a possible questionnaire flow for a respondent could have been as follows:

- Task 1 (indicate best and worst concepts)
- Task 2 (indicate best and worst concepts)
- Task 3 (indicate best and worst concepts)
- Task 4 (indicate best and worst concepts)
- Task 5 (indicate best and worst concepts)
- Probe if any not-selected levels are unacceptable (respondent indicates "Las Vegas")
- Task 6 (indicate best and worst concepts)
- Probe if any not-selected levels are unacceptable (respondent indicates "3 night stay")
- Task 7 (indicate best and worst concepts)
- Probe if any not-selected levels are unacceptable (respondent indicates "Western Europe")
- Task 8 (indicate best and worst concepts)
- Probe if any not-selected levels are unacceptable (respondent indicates "No more are unacceptable)

- Task 9 (indicate best and worst concepts)
- Task 10 (indicate best and worst concepts)
- Task 11 (indicate best and worst concepts)
- Task 12 (indicate best and worst concepts)

This particular respondent indicated that three levels in the study were unacceptable. Four additional questions (four unacceptable level probes) were asked of the respondent beyond what would have been done in a standard B-W CBC questionnaire.

For utility estimation, we coded each choice task seen by respondents as two tasks (choice of best concept, and choice of worst concept, similar to what is done with best/worst scaling). For the best task coding, we actually included five total concepts: the four concepts seen by the respondent, plus a "None" concept (and associated dummy-coded None column). The None concept is never chosen in these "best" tasks. The worst task coding simply inverted the design matrix for the four (non-None) concepts from the best task. And, any of the unacceptable concepts (concepts that included an unacceptable level) that had been modified by the adaptive algorithm were coded as additional paired-comparison tasks versus the None concept, where the None concept was indicated as selected each time. So, for this example respondent, if 14 concepts had been found to include levels "Las Vegas," "3 night stay," or "Western Europe," 14 additional tasks would be appended to this respondent's data (for a total of 24+14 = 38 coded tasks). This procedure led to estimates for unacceptable levels that were generally inferior to acceptable levels within the same attribute, but that were not assumed to be absolute unacceptable (negative infinity). The Bayesian shrinkage within CBC/HB estimation kept these unacceptable levels from drifting toward negative infinity, and also used information regarding the sample means to inform the relative (negative) utility of the unacceptable levels.

For the respondents that completed the A-DCM survey, 71% of them indicated that at least one level was unacceptable. The most number of unacceptable levels indicated by any one of these respondents was 5. We will discuss the unacceptable levels in more detail later, and compare them with those in the Fixed Designs.

B) Satisfaction of Respondents

Respondents showed some preference for the adaptive DCM survey versus the traditional Fixed Conjoint. In the fixed designs, 9-10% of respondents dropped out during the conjoint tasks. In contrast only 5% of those in the Adaptive DCM dropped out.

	Fixed Low Overlap	Fixed High Overlap	Adaptive DCM
Median Time to Complete	13.5 min	13.3 min	10.4 min
% Respondents started survey	100%	100%	100%
% Respondents finished conjoint	91%	90%	95%
% Respondents completed additional ratings per level	90%	89%	N/A
% Respondents finished survey	89%	89%	95%

The fixed cells (low and high overlap) took longer to complete because we asked a series of rating scale questions about the conjoint attributes after all the conjoint scenarios. We will discuss these rating scale questions in more detail later. The initial analyses do not use them. It is interesting to note from the table above that almost no respondents dropped out during the rating scale exercise.

As the table below shows, respondents also found the adaptive survey less monotonous and boring. So the adaptive DCM is preferred by respondents in many ways, though they were equally likely to express desire to take future surveys of all three types.

	Low Overlap	High Overlap	Adaptive DCM
Monotonous and boring	27%	25%	17%*
Vacation packages were realistic	80%	72%*	79%
Interested in taking another survey like this in the future	84%	82%	83%
Survey format made it easy for me to give realistic answers	80%	77%	81%
Survey format made me want to slow down and make careful choices	83%	80%	81%

* Significantly lower than expected

4. Basic Comparative Results of 3 Test Cells

The same coding was used for all 3 cells, except the Fixed Cells did not have the synthetic None option. The latter was added to the Adaptive DCM to account for unacceptable levels. We ran HB estimation for each of the three test cells. We then tuned the scale factor separately within each cell using the 3 calibration tasks (which were not used for estimating utilities). These same 3 calibration tasks are used as holdout tasks for predicting hit-rate.

All three designs gave comparable results.

	Total Hit Rate	Holdout 1	Holdout 2	Holdout 3
Adaptive DCM	41.6%	**42.9%**	38.0%	43.9%
Fixed Cell A (Low Overlap)	**43.0%**	40.3%	**41.3%**	**47.4%**
Fixed Cell B (High Overlap)	38.9%	36.5%	39.7%	40.4%

The low overlap cell shows a pretty nice advantage predicting holdout task 3. Had that been our only holdout, we would likely think it is the winner. But Adaptive DCM does better in

Holdout Task1. One of the morals of the story is that you should always test more than one holdout.

The high overlap cell shows some signs of being weakest here, as it comes in last two out of three times. It has only one second place finish, barely beating the adaptive cell in holdout 2.

When it comes to predicting the out of sample tasks (12 tasks x 6 alternatives), there is even less difference between the cells. But this makes deciding on a winner even more confusing, as high overlap has a very small advantage (certainly not significant).

	R^2	MAE
Adaptive Conjoint	.776	3.2%
Fixed Cell A (Low Overlap)	.797	3.2%
Fixed Cell B (High Overlap)	**.800**	**3.1%**

The fixed cells are nearly identical here, with a very slight advantage over Adaptive DCM in R^2. High Overlap cell is certainly not the worst.

These results all assume we asked Best and Worst choices for the four alternatives. We thought it worthwhile to compute what might have happened had we asked only for the best choice.

Best Choice Only	Hit Rate	R^2	MAE
Adaptive DCM	42.4%	0.747	3.6%
Fixed Cell A (Low Overlap)	**44.5%**	**0.786**	**3.5%**
Fixed Cell B (High Overlap)	37.6%	0.769	3.8%

When using only the Best choices to estimate utilities, the low overlap cell seems the likely winner. At the very least the low overlap cell consistently beats the high overlap cell. The hit-rate is almost 7% higher, and this is consistent across all 3 holdout tasks. The MAE and R^2 are also better than the high overlap.

One very interesting finding is that for the adaptive and low overlap cells, the Best Only model actually has a higher hit rate than asking Best and Worst. For the low overlap cell this is true across all 3 holdouts. The reverse is true for the high overlap cell – adding the worst improved the hit rate. So adding the worst choice helps the high overlap cell much more than the low overlap cell, closing the gap between the two designs.

		Total Hit Rate	Holdout 1	Holdout 2	Holdout 3
Low Overlap	Best Only	**44.5%**	**43.0%**	**42.5%**	**47.9%**
	Best + Worst	43.0%	40.3%	41.3%	47.4%
High Overlap	Best Only	37.6%	35.2%	37.7%	40.0%
	Best + Worst	38.9%	36.5%	39.7%	40.4%

Note that adding the worst choice helps out of sample R^2 and MAE for all three test cells, though it helps high overlap cell the most. So while we question the benefit of asking the worst choice for low overlap designs, this study gives some evidence that asking the worst choice could provide benefit with high overlap designs.

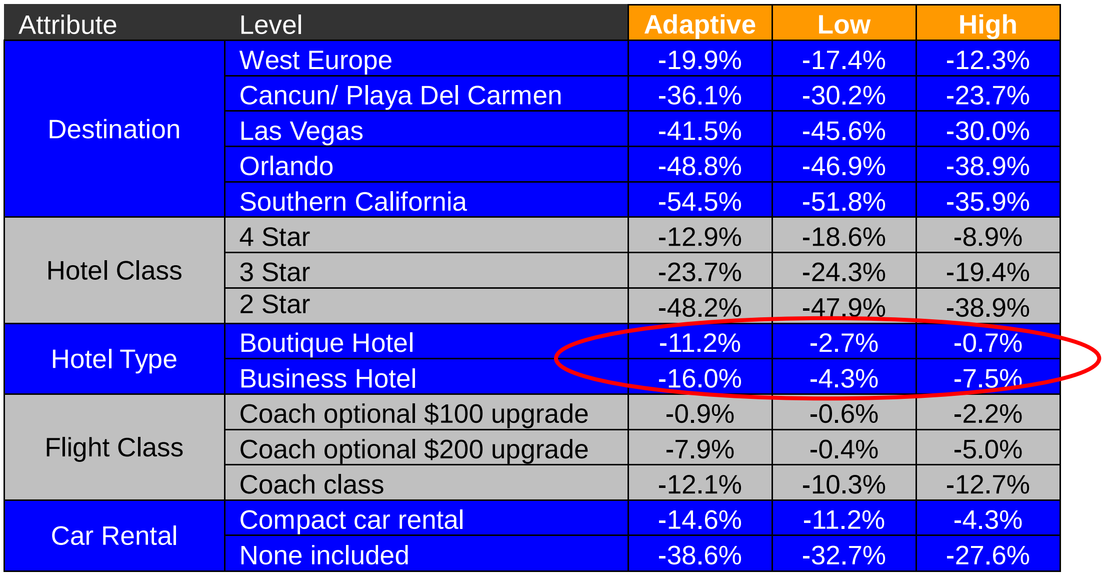
The difference in how the low overlap and high overlap cell react to the worst choice led us to dig deeper into how these different designs compare with one another.

5. Comparing Predictions of Test Cells

All 3 cells create similar predictions for the 12 tasks in the holdout cell. The correlation between the three cells across the 72 predicted shares (12 tasks x 6 alternatives) is .93 - .95. That said there are 14 alternatives where the predicted shares differ by more than 5%. So in some specific cases one will see a somewhat noticeable difference in share.

More systematically, we can do something like comparing the mean utilities, but using a choice simulator. Specifically, we applied a similar but more rigorous analysis, estimating the deviations from a base case. These deviations are estimated by comparing two vacation packages. One is a base case. In this case, the base case package is the near optimal vacation package: a 7 night vacation to Hawaii at a 5 star resort, with SUV car rental and 1st class flight for $1950. We compare this base vacation package with the exact same package, except for one change. For instance we might change the destination to Western Europe. This gives us a difference in share between the base case and the new package. That difference is the deviation. We repeat this pairwise test of base vs. base + 1 change for each level not in the base case. The result is a set of deviations that show how much better or worst each level is, given the base case.

The deviations are highly correlated with one another, at .97 for all 3 cells with each other. The high overlap cell showed less deviation than the other cells. The exception to this was Hotel Type and Flight Class, where the High Overlap cell showed more deviation than the Low Overlap Cell.



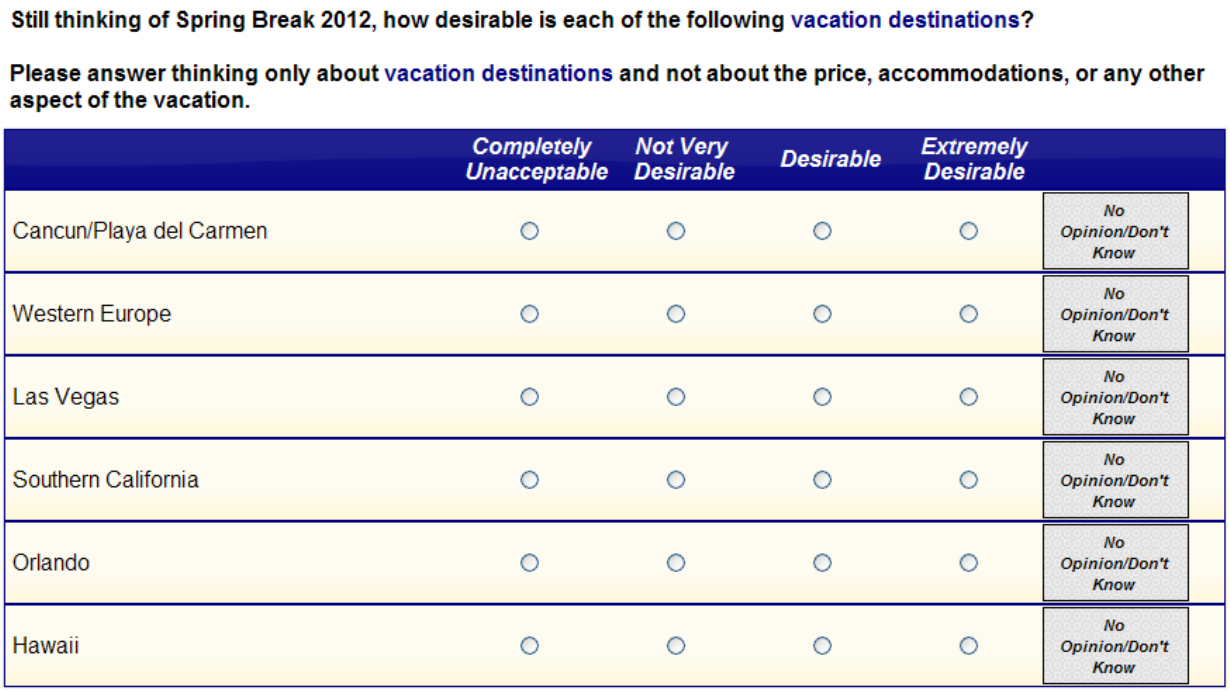
Attribute	Level	Adaptive	Low	High
Destination	West Europe	-19.9%	-17.4%	-12.3%
	Cancun/ Playa Del Carmen	-36.1%	-30.2%	-23.7%
	Las Vegas	-41.5%	-45.6%	-30.0%
	Orlando	-48.8%	-46.9%	-38.9%
	Southern California	-54.5%	-51.8%	-35.9%
Hotel Class	4 Star	-12.9%	-18.6%	-8.9%
	3 Star	-23.7%	-24.3%	-19.4%
	2 Star	-48.2%	-47.9%	-38.9%
Hotel Type	Boutique Hotel	-11.2%	-2.7%	-0.7%
	Business Hotel	-16.0%	-4.3%	-7.5%
Flight Class	Coach optional $100 upgrade	-0.9%	-0.6%	-2.2%
	Coach optional $200 upgrade	-7.9%	-0.4%	-5.0%
	Coach class	-12.1%	-10.3%	-12.7%
Car Rental	Compact car rental	-14.6%	-11.2%	-4.3%
	None included	-38.6%	-32.7%	-27.6%

Hotel Type is the most interesting attribute. In the adaptive version, hotel type shows much more deviation from the base type (resort). Later we will show why we think the adaptive deviations are more accurate than the low or high overlap cells. Comparing the adaptive vs. the low overlap cell, one can see that this is really the only attribute where they vary significantly. Another interesting difference can be found when we look at low overlap vs. high overlap. For Hotel Type, high overlap shows a larger negative difference than low overlap, while for every other attribute the high overlap cell shows much smaller negative deviations.

So while the cells are comparable in many ways, we certainly see some differences, especially on Hotel Type, and for the lower deviations in the High Overlap cell.

6. More Diagnostics: Comparing Utilities with Ratings

For the low overlap and high overlap cells, we asked respondents to rate the levels of each attribute, as shown below for Destination.



Still thinking of Spring Break 2012, how desirable is each of the following vacation destinations?

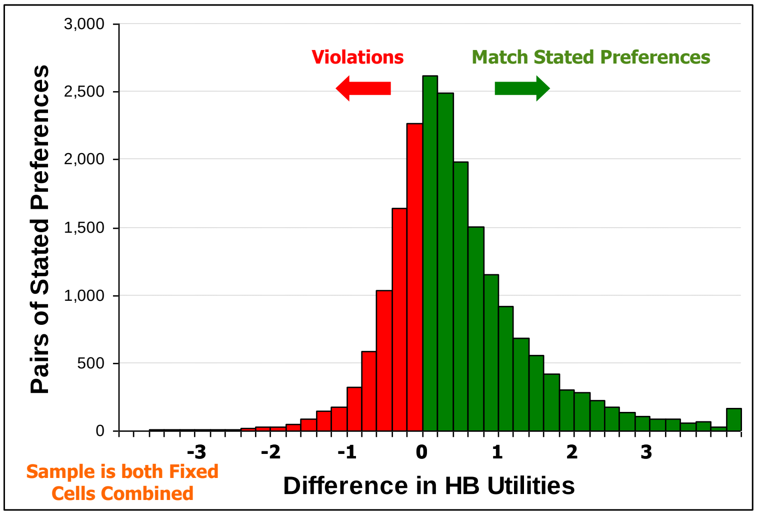
Please answer thinking only about vacation destinations and not about the price, accommodations, or any other aspect of the vacation.

	Completely Unacceptable	*Not Very Desirable*	*Desirable*	*Extremely Desirable*	
Cancun/Playa del Carmen	○	○	○	○	*No Opinion/Don't Know*
Western Europe	○	○	○	○	*No Opinion/Don't Know*
Las Vegas	○	○	○	○	*No Opinion/Don't Know*
Southern California	○	○	○	○	*No Opinion/Don't Know*
Orlando	○	○	○	○	*No Opinion/Don't Know*
Hawaii	○	○	○	○	*No Opinion/Don't Know*

Later, we will discuss the completely unacceptable responses specifically. But what we can do is treat this information as sets of pairwise ordinal relations. So if a respondent said Orlando was Desirable while Las Vegas was Completely Unacceptable, we can infer that the utility of Orlando should be higher than Las Vegas for this respondent.

When we look at all the pairs and compare them with the utilities, we get many reversals. In fact, there are 17,173 pairs across the 800 respondents in the Low and High Overlap cells. About 29% of these pairs are reversals, with 23% of them having significant reversals (more than .1 difference).

The histogram below shows for each pair where Level A beats Level B, the difference in HB utilities. We would like to see that all the differences are positive, meaning the utility difference matches the ordinality of the stated rating.



We can break this comparison down in more detail, looking at the reversals by attribute and level. In this case we limit the reversals so that Higher Rated – Lower Rated < -.1. This eliminates the very small reversals that have little meaning.

	% Pair Violations (<-.1) in Utilities Among All Stated Pairs			Counts of Utility Violations		Respondents with 1+ Violations	
	Low Overlap	High Overlap	Total	Low Overlap	High Overlap	Low Overlap	High Overlap
Destination	18.1%	**21.4%**	19.7%	699	785	242	263
Nights	Not Applicable						
Hotel Stars	**19.8%**	17.0%	18.4%	352	295	185	169
Hotel Type	**35.9%**	28.6%	32.3%	279	218	188	152
Flight Class	**35.7%**	27.8%	31.8%	523	388	254	213
Rental Car	**27.5%**	19.1%	23.3%	240	163	171	131

At Maritz, we (Lattery and colleagues) have conducted this kind of analysis over many data sets. What we typically find is that there are more reversals in less important attributes. HB estimation tends to get the more important attributes right, while less important attributes have more room to fluctuate. One can see in the table above that Destination and Hotel Stars (the more important attributes) have far fewer reversals than Hotel Type and Flight Class (least important attributes).

When we add overlap in the Design to more important attributes, this gives us more information about the less important attributes. You can see in the High Overlap design that there are fewer reversals, especially for Hotel Type, Flight Class, and Rental Car. Hotel type and Car Rental had minimal overlap in both designs. But by adding overlap to more important attributes we get better information about less important attributes. This is the information we need to supplement HB, since HB is more likely to produce reversals in less important attributes.

One can clearly see that the high Overlap cell has fewer reversals. This is true for every attribute, except Destination where Low Overlap has slightly fewer reversals. But overall there are more reversals for the Low Overlap cell, and this is even more prominent when we make the cutoff stronger than -1. As we count only the reversals that are bigger, we see that the Low Overlap cell has even more reversals

Minimum Reversal	Low Overlap	High Overlap	High/ Low
0	28.7%	28.9%	100.9%
-0.1	23.9%	21.9%	91.7%
-0.2	19.7%	15.8%	80.2%
-0.3	15.8%	11.0%	69.6%
-0.4	12.5%	7.6%	61.3%
-0.5	9.5%	5.1%	53.5%
-0.6	7.3%	3.5%	47.7%

In fact at a 0 cutoff the cells are equal, with the High Overlap cell having slightly more reversals. But as we make the reversals stronger, we see more reversals in the Low Overlap cell. So the difference between these two cells widens, meaning we have more significant reversals in the High Overlap cell.

It seems the overlap on more important attributes gives us better information within an attribute (fewer reversals). Now if the High Overlap cell also gave better information about the relative importance of attributes then we would expect the High Overlap cell to have better fit. But, if anything, the low overlap cell has better fit. So we conclude that the High Overlap cell gives us less accurate information about attribute importance. This conclusion is further supported by the fact that the deviations in the Low Overlap cell looked different from the High Overlap and Adaptive cells. The consistency of the deviation found in the Low Overlap and Adaptive cells suggests the High Overlap cell is the less accurate odd man out. *So our working hypothesis is that the High Overlap cell gives better information within attributes by sacrificing information about the relative importance across attributes.*

7. Developing a Better Model by Including the Ordinality of Ratings

Adaptive DCM used additional information about unacceptable levels in its analysis. We can do something similar with the Fixed Design cells. In this case we develop models that include respondent level information from the ratings. Lattery (2009) shows how this can be done in HB by adding simulated tasks of two alternatives, where the higher rated alternative (specified to vary only on one attribute) beats the lower rated alternative. Another method is to apply Expectation-Maximization as described by Lattery (2007). This method estimates conjoint utilities for each respondent one at a time, and allows each respondent’s utilities to be completely customized with unique constraints.

The best model that emerged was when we applied EM estimation with respondent level constraints in the low overlap cell. This model is shown below, along with the other three cells previously discussed.

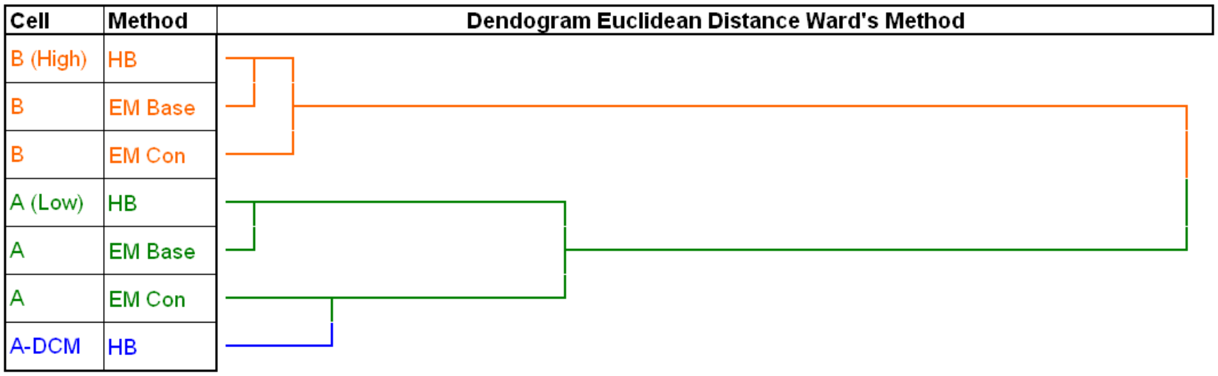
	Total Hit Rate	Task1	Task 2	Task 3	R^2
HB Adaptive DCM	41.6%	42.9%	38.0%	43.9%	.776
HB High Overlap	38.9%	36.5%	39.7%	40.4%	.800
HB Low Overlap	43.0%	40.3%	41.3%	47.4%	.797
EM Low Overlap w/Constraints	44.7%	42.8%	43.5%	47.9%	.837

The EM model with respondent constraints has the highest hit rate, except for Task 1 where Adaptive DCM had a .1% advantage (about 1 person). The out of sample R^2 shows a significant increase, going from .80 to .84.

Perhaps even more important than the improved hit rate and R^2 is the deviations we get. Recall that the adaptive and low overlap cell were most similar to one another. Their big difference was on Hotel Type. Now when we add the constraints to the Low overlap cell and estimate via EM, we get results that are most similar to the adaptive cell.

Attribute	Level	Adaptive	Low	High	EM (Low)
Destination	West Europe	-19.9%	-17.4%	-12.3%	-21.9%
	Cancun/ Playa Del Carmen	-36.1%	-30.2%	-23.7%	-32.9%
	Las Vegas	-41.5%	-45.6%	-30.0%	-43.3%
	Orlando	-48.8%	-46.9%	-38.9%	-46.4%
	Southern California	-54.5%	-51.8%	-35.9%	-48.9%
Hotel Class	4 Star	-12.9%	-18.6%	-8.9%	-17.5%
	3 Star	-23.7%	-24.3%	-19.4%	-27.7%
	2 Star	-48.2%	-47.9%	-38.9%	-51.5%
Hotel Type	Boutique Hotel	-11.2%	-2.7%	-0.7%	-6.2%
	Business Hotel	-16.0%	-4.3%	-7.5%	-16.2%
Flight Class	Coach optional $100 upgrade	-0.9%	-0.6%	-2.2%	-1.9%
	Coach optional $200 upgrade	-7.9%	-0.4%	-5.0%	-8.9%
	Coach class	-12.1%	-10.3%	-12.7%	-11.5%
Car Rental	Compact car rental	-14.6%	-11.2%	-4.3%	-12.9%
	None included	-38.6%	-32.7%	-27.6%	-36.3%

We also ran an EM Model with constraints for the High Overlap cell, but it showed far less difference from the HB model without respondent level constraints. We also ran EM models without any respondent constraints, and the results were nearly identical to their HB counterparts. In fact the errors for out of sample prediction were correlated .97. The following dendogram shows the various models we ran, and groups them in terms of similarity.



Dendogram Euclidean Distance Ward's Method

Cell	Method
B (High)	HB
B	EM Base
B	EM Con
A (Low)	HB
A	EM Base
A	EM Con
A-DCM	HB

The EM Base models without constraints were most similar to their HB counterparts. So estimation method was less relevant than adding constraints. Perhaps the most interesting aspect of this dendogram is that the EM Constrained model for the Low overlap cell (which has the best fit) is most similar to the Adaptive DCM cell. We think the adaptive DCM with its additional analysis of unacceptables probably did a better job in capturing the story of the level deviations than Fixed Designs without any rating information. Of course, we think the EM Constrained Low Overlap model using respondent ratings captured the story even more accurately since it has the best hit rate and out-of-sample prediction.

One of the conclusions from this study that Maritz has confirmed in many others is that without additional information from ratings or unacceptables, HB estimation tends to be inaccurate on attributes that are less important, such as Hotel Type. So these less important attributes may have an unclear resolution on the relative value of their levels Lattery (2009) illustrates this point in more detail, adding respondent level constraints via both EM and HB to get better resolution on levels within an attribute and thereby tell a more accurate story.

8. Non-Compensatory Elimination of Unacceptable Levels

Adaptive DCM used an adaptive probing method to elicit unacceptable levels. The fixed cells asked respondents directly via a simple rating scale. When we looked at this data in more detail, it appears that respondents overstate unacceptable levels.

The raw percentages in the table below show how many respondents said a level was unacceptable on the rating scale. We then compare what they said with their actual choices in the 12 tasks. If they choose an alternative with an unacceptable level, and there was another alternative in the task without any unacceptable levels, then that level is not really unacceptable. This falsifies the unacceptability of that level. If there are no falsifications across the 12 tasks, then we have confirmation. It is worth noting that confirmation does not guarantee the level will always be unacceptable, just that it is consistent (not falsified) with their observed responses.

	Raw		Confirmed	
	Low	High	Low	High
Hawaii	0.5%	2.0%	0.0%	1.0%
Las Vegas	8.6%	6.0%	6.6%	4.2%
Orlando	7.9%	4.5%	6.4%	3.0%
Southern California	5.4%	5.0%	3.7%	3.0%
Western Europe	9.6%	7.7%	7.1%	4.2%
Cancun/Playa Del Carmen	9.1%	8.7%	7.1%	6.9%
3 nights	**15.0%**	**14.1%**	**2.9%**	**1.0%**
4 nights	3.9%	3.2%	0.0%	0.2%
5 nights	1.0%	1.0%	0.0%	0.0%
7 nights	2.9%	2.0%	1.0%	0.7%
Moderate (2 star)	**26.5%**	**22.3%**	**8.1%**	**7.2%**
Upscale (3 star)	2.9%	2.5%	0.2%	0.0%
Deluxe (4 star)	0.7%	1.0%	0.0%	0.0%
Luxury (5 star)	1.2%	2.2%	0.0%	0.0%
Business (with meeting/business services)	6.4%	8.2%	0.2%	0.7%
Resort (usually with spa, golf, etc.)	1.0%	0.5%	0.0%	0.0%
Boutique (with distinct style/character)	2.2%	1.2%	0.0%	0.0%
Coach class	6.1%	2.0%	1.0%	0.2%
Coach with optional $100 upgrade to business/1st class	1.5%	0.2%	0.2%	0.0%
Coach with optional $200 upgrade to business/1st class	5.4%	5.5%	0.5%	0.0%
Business/1st class	2.2%	2.0%	0.0%	0.0%
None included	**21.4%**	**17.4%**	**2.2%**	**1.7%**
Compact car rental	3.4%	1.0%	0.0%	0.0%
Full-Size/SUV rental	2.0%	1.2%	0.0%	0.2%

For instance 2 star hotels are clearly undesirable, and 26.5% of respondents said they were completely unacceptable. But less than 1/3 of those respondents were consistent in their choices – more than 2/3 never picked a 2 star hotel when another alternative was available without unacceptable levels. So in this context, when a respondent says a level is unacceptable, it tends to mean the level is highly undesirable, but if there are other good things in the alternative, the level is not really unacceptable.

One hypothesis is that respondents may apply non-compensatory decisions as a simplifying strategy to reduce larger number of choices to a consideration set. In a real world setting, respondents who think a 2 star hotel is completely unacceptable would likely find some choices with 3 or more stars. In our case, with four vacation choices, they compromised. So a respondent might say a 2 star hotel is unacceptable, but then they make an exception when the 2 star hotel is in Hawaii with a really cheap price, and the other 3 alternatives are more expensive.

The confirmed percentages are similar across the low and high overlap cells. These confirmed percentages appear truly unacceptable – screening out those alternatives with those unacceptable levels. Given the random sampling, we expect similar percentages would also apply to the Adaptive DCM cell. When we look at how many respondents think levels are unacceptable we see they are quite a bit higher than the confirmed percentages.

	Raw			Confirmed	
	Low	High	Adaptive	Low	High
Hawaii	0.5%	2.0%	4.5%	0.0%	1.0%
Las Vegas	8.6%	6.0%	9.9%	6.6%	4.2%
Orlando	7.9%	4.5%	11.7%	6.4%	3.0%
Southern California	5.4%	5.0%	**15.1%**	3.7%	3.0%
Western Europe	9.6%	7.7%	**16.4%**	7.1%	4.2%
Cancun/Playa Del Carmen	9.1%	8.7%	**17.6%**	7.1%	6.9%
3 nights	**15.0%**	**14.1%**	11.2%	**2.9%**	**1.0%**
4 nights	3.9%	3.2%	2.7%	0.0%	0.2%
5 nights	1.0%	1.0%	1.5%	0.0%	0.0%
7 nights	2.9%	2.0%	5.0%	1.0%	0.7%
Moderate (2 star)	**26.5%**	**22.3%**	**16.6%**	**8.1%**	**7.2%**
Upscale (3 star)	2.9%	2.5%	1.0%	0.2%	0.0%
Deluxe (4 star)	0.7%	1.0%	0.2%	0.0%	0.0%
Luxury (5 star)	1.2%	2.2%	0.0%	0.0%	0.0%
Business	6.4%	8.2%	3.2%	0.2%	0.7%
Resort	1.0%	0.5%	0.7%	0.0%	0.0%
Boutique	2.2%	1.2%	1.0%	0.0%	0.0%
Coach class	6.1%	2.0%	2.7%	1.0%	0.2%
Coach with optional $100 upgrade to business/1st class	1.5%	0.2%	1.7%	0.2%	0.0%
Coach with optional $200 upgrade to business/1st class	5.4%	5.5%	3.7%	0.5%	0.0%
Business/1st class	2.2%	2.0%	2.0%	0.0%	0.0%
None included	**21.4%**	**17.4%**	8.2%	**2.2%**	**1.7%**
Compact car rental	3.4%	1.0%	1.2%	0.0%	0.0%
Full-Size/SUV rental	2.0%	1.2%	0.2%	0.0%	0.2%

The Adaptive percentages are typically higher than what we see in the confirmed cells. So while the numbers look better than the raw ratings, they are still too high. We also see that for the destination levels, the unacceptable rates in Adaptive are actually higher than the raw. One reason for this is that the confirmed unacceptables are not too much lower than the raw. In general, respondents might compromise on hotel stars, but they held steadfast on their raw unacceptable levels when it came to destinations. But the values in red are significantly higher than even the raw ratings. It was pointed out that the number of levels impacts the unacceptable levels in the Adaptive probing method. Since Destination has more levels, the levels appear less frequently in the initial 5 tasks, making them more likely to appear in the initial list of potential unacceptables. In general this is a problem for adaptive probing, as the number of levels will impact the chance of a level's appearing on the candidate list of unacceptables.

That said, the lack of accurately capturing unacceptables is not a serious problem. The modeling we have run only forces the utilities of these levels downward. In the case of the Fixed EM cells, the unacceptables were simply constrained to be lower than other levels. In the adaptive case, the None option beat alternatives with the unacceptable levels.

It is however vitally important to recognize that any method of analysis incorporating information about unacceptable levels recognize the very real possibility that respondents seriously overstate the unacceptability of levels. Scoring these unacceptable levels with very large negative utilities may actually be detrimental. At the very least one needs some additional checks (like conforming tasks) before doing this.

Having done dozens of conjoint studies with additional attribute rating information, Maritz can attest to the pattern seen in this study. There may be one or two attributes where stated unacceptables are confirmed in how respondents choose, but many other attributes where "unacceptable" really means "less desirable".

9. Conclusion

Hit-rate and out of sample prediction between the three different designs were very similar. No design stood out as a clear winner. Adding the Worst Choice question per task did not help much with overall fit. In general, we are unconvinced that adding the worst choice is worth much. If the researcher is going to ask the respondent to do a bit more effort, perhaps a couple more Best-Only choice tasks would be better than Worst Choices accompanying every task. That remains an interesting avenue for future research.

In this study, the high overlap design stood out a bit from the other cells. The higher overlap gave us better information about levels within attributes (fewer reversals), but probably sacrificed some information about attribute importance. The high overlap cell also had the most to gain from adding the worst choice. Asking both best and worst improved the model, while in the other cells it had no real impact. Finally, the high overlap cell showed different deviations from the low and adaptive cells. Unfortunately we did not confirm the results from Chrzan, Zepp, White (2010) that suggest high overlap on more attributes will improve overall fit. That said, this was only one study, and may be an anomaly.

One of the conclusions from this study that we have confirmed in many others is that without additional information from ratings or unacceptables, HB estimation tends to produce many reversals in utilities, especially on attributes that are less important, such as Hotel Type in this study. This means some of the less important attributes may have an unclear resolution on the relative value of their levels. Lattery (2009) illustrates this point in more detail. Using EM estimation with respondent level constraints eliminates reversals by adding within attribute information, and gives the best model with low overlap design. Interestingly, the adaptive DCM was the most similar to this model in its story.

Finally, we confirm again that respondents overstate unacceptable levels. Adaptive probing might help reduce overstatement but still showed more unacceptable levels than expected. It is important than any method of analysis that incorporates information about unacceptable levels recognize the very real possibility that respondents seriously overstate the unacceptables. Scoring these unacceptable levels with overly large negative utilities may actually be detrimental. At the very least one needs some additional checks (like consistency across tasks). In this study we basically treated unacceptables as more negative than other levels that were not unacceptable.

REFERENCES

Chrzan, Keith, John Zepp and Joseph White: The Success of Choice-Based Conjoint Designs among Respondents Making Lexicographic Choices. 2010. *Proceedings of the Sawtooth Software Conference*, pp 19-36.

Johnson, Richard and Bryan Orme: A New Approach to Adaptive CBC. 2007. *Proceedings of the Sawtooth Software Conference*, pp 85-110.

Lattery, Kevin: EM CBC: A New Framework For Deriving Individual Conjoint Utilities by Estimating Responses to Unobserved Tasks via Expectation-Maximization (EM). 2007. *Proceedings of the Sawtooth Software Conference*, pp 127-138.

Lattery, Kevin: Coupling Stated Preferences with Conjoint Tasks to Better Estimate Individual Level Utilities. 2009. *Proceedings of the Sawtooth Software Conference*, pp 171-184.

DISCUSSION OF LATTERY AND ORME

RICH JOHNSON
SAWTOOTH SOFTWARE

Kevin and Bryan have done interesting and useful work. It's disappointing that the adaptive treatment wasn't clearly superior, but several iterations are often required in methods development. Their findings point the way to what I think could become a useful new approach. They also confirmed a couple of points that are of more general importance. I'd like to concentrate on two of those points, and then suggest what they might look at in their next study.

First, we need to be wary about "Unacceptables." One of their most interesting slides shows the percentage of respondents who declared each level to be "unacceptable," and also the percentage of respondents actually accepting an unacceptable level in a chosen alternative. Averaging across all levels for respondents in Cells A and B, levels had average likelihood of about 6% of being declared unacceptable. But about 70% of the times a respondent declared a level to be unacceptable, it wasn't really, because that respondent actually chose an alternative containing that same level. I believe that when a respondent declares a level to be unacceptable, he is really just telling us that he would rather have something else. It would be a mistake to take such judgments literally. To their credit, Kevin and Bryan didn't do that; instead, they used the more conservative approach of treating that judgment as an indication that the respondent would prefer concepts not including that level.

Second, I think it's a good idea to merge choice data with rating scale data. Kevin and Bryan used each respondent's desirability ratings in the two fixed cells to constrain his order of partworths within each attribute, and this additional information produced an improvement in hit rates and share prediction accuracy for holdout respondents. We found a similar result with ACA several years ago (at least for the hit rates), which led us to provide that capability in the default settings for HB estimation offered within ACA/HB, and it's also an advanced option in ACBC. I believe that the respondent comes to the interview with a sense of the relative desirability of each level within an attribute. This information can be obtained quickly and with little pain to the respondent, which also serves as a good educational warm-up task regarding the attribute levels. This information is so easy to get, and contributes so much value, that it seems wasteful not to get it.

Now, my suggestion about what to try next would be an interview which starts with a series of rating grids like those used in their "fixed" cells (though without "unacceptable" as a descriptor). That could be followed by a section identical to the current A-DCM module. The analysis might be done using HB or EM, applying within-individual constraints compatible with the information from the first section. I think their results suggest that such a procedure should work well.

I want to end by thinking Kevin and Bryan for a thought-provoking presentation. This was good work, and it points us in a useful direction for future improvement.

Being Creative in the Design: Performance of Hierarchical Bayes with Sparse Information Matrix.

Jay Weiner
Ipsos Media-CT
Marcos Sanches
Ipsos Reid

1. Abstract

The usual goal of the researcher when designing a Discrete Choice exercise is to achieve a high efficiency design. In some cases, however, the reality is more complex and following book recipes just will not work. We will present one such a case in which design and analysis will provide interesting insight on how to get out of the box on handling unusual situations.

The key feature of our problem is that in the face of the impossibility of generating a full profile design we decided to combine attributes and present many concepts in a single board. The results provide some interesting insights on two areas not explored a lot in discrete choice literature: consequences of having many concepts per board and the use of simpler discrete choice models.

2. Introduction

Here we will present a single case, its design and analyses. To keep the confidentiality of our client we will not show brand or product names, yet the results are real. The product is a fast moving package good, with a short purchase cycle. Variety seeking is common in this category.

The current market is composed of 14 main brands (B1, B2… B14), each brand has one, two or three sizes (specific to the brand) and can be offered in two different types of package. Additionally each brand has both regular and sales prices. The client brand (B9), with two current SKUs in the market, wants to launch several new SKUs (8 possible new alternatives). The client also wants to have an idea of price sensitivity and sales. Table 1 describes the scenario with its attributes and levels.

The matrix in Table 1 shows that the configuration of products in the market is very irregular. Sizes are specific to Brands, and Package Type is specific to Brand and Size. Besides, we still have two possible Prices (Regular and Sales) for each entry in Table 1, which must be tested. These prices further add to the complexity since they are again specific to Brand and Size.

The business objective is primarily to test the products in bold in Table 1 (Brand “B9”, Package Type “Bag” – they currently have only Package Type “Tin” – and different sizes) and understand how they behave in the market. The client is also interested in some forecast for the new product and simulation of performance against competitor’s prices as well as insights about possible cannibalization of existing products.

Table 1. Combinations of Brand and Package (T=Tin; B=Bag) currently in the market. Package Type Bag in bold for Brand 9 are new products, currently not in the market.

Size/Brand	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
100g														B
120g									**B**					
155g									**B**					
240g									**B**					
280g									**B**					
305g							T		**B**					T
328g	T								T					
349g													T	
350g									**B**					
434g				B					**B**		B			
450g									**B**			B		
454g		B			B			B		B				B
600g			T											
680g							B							
900g										T				
915g									T					
926g						T								
930g	T												T	
990g			T											

3. Design

The first alternative was to generate a full profile design. Assuming that Brand, Size and Package are factors, Table 1 shows that such a full profile design would mostly result in products not in the market and therefore not of interest. Many products would be strange, with prices not appropriated for their sizes or brand. People familiar with this market would find strange certain combinations of Brand, Size and Package. The attributes Size and Brand also have way too many levels considering the sample size we were willing to use. Finally, there is no easy way to represent the Price because Regular and Sales prices depend on Brand, Size and Package.

The second alternative was representing the set of configurations in Table 1 as it is, by using an Alternative Specific Design (or a full profile design with prohibitions) where Size would be specific to Brand and Package Type specific to Brand and Size. But such a design would have too many prohibitions and certainly would not be efficient. We did not feel comfortable with any of these alternatives.

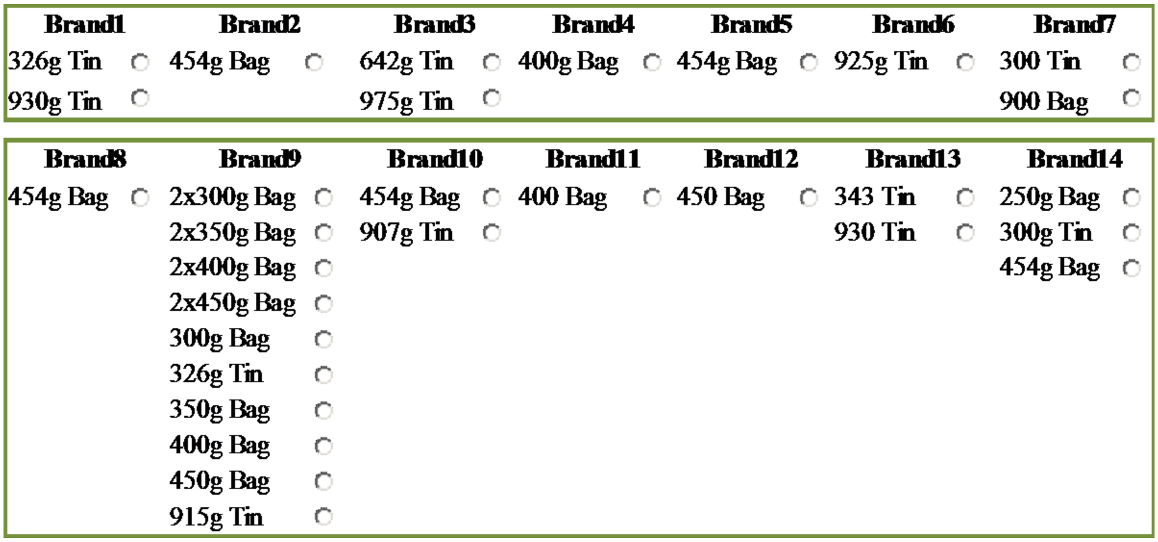
If the client just wanted to understand which of the new products (in bold in Table 1) they should launch then a methodology that would come to mind would be a monadic or sequential monadic test, where we would just ask respondents to evaluate each of the new products, which are the alternatives the client wants to launch in the market. That is not enough, though, because the client is looking for Discrete Choice outputs. They want to have an idea of Market Share for their new product, cannibalization and also be able to simulate scenarios where competitors' products go on promotion. The important piece of information here is that to address this business objective we don't really need to have utilities/ importance for Brand, Size and Package Type as individual attributes. So another option is to collapse Brand, Size and Package Type into a single factor: SKU. And now we have SKU with 30 levels, most of them with two Price points: Regular and Sales that are specific to the SKU. That is the route we decided to take.

The new SKU attribute has 30 possible levels. In order to have a good standard error for the utilities we need a large sample, or we need to show many concepts per task, or many tasks per respondent. Sample size is always a serious limitation since increasing sample is associated with increased costs. Increasing the number of concepts per task seemed like an interesting idea because the real market actually offers many options. We realized that in order to have a realistic evaluation of the concept we should have all the concepts in the same task, which would mean 30 concepts per task. Although studies have shown that more concepts per task are better than fewer in providing statistical information and avoiding biases (Pinnell & Englert 1997) and in increasing design efficiency (Zwerina, K., Huber, J & Kuhfeld, W 1996), the number of concepts should be limited by the complexity of the attributes (Cunningham, C. E., et. al. 2004). We think that being in the fast moving package good category with short purchase cycle and having only 4 attributes make our concept quite simple. People are routinely making this sort of choice among many options when they are in front of the shelf in a grocery store. Furthermore, one could argue that presenting all the SKUs makes the choice task easier when the respondent goes further into the exercise, since the concept order does not need to change from task to task and the SKU attribute will not vary from task to task, the only thing varying being the Price. If we presented only, say, 10 concepts per task then the respondent would see completely different SKUs and prices when moving from one task to the following. Additionally, we used formatting to further simplify the respondent's work, by showing only 14 instead of 30 columns, tying each column to a brand so that each brand had its different sizes under the same column in the board (see Figure 1). Concluding, not only there is an advantage in term of realism when all concepts are shown, but we could also have an advantage in terms of simplicity. With this example we argue that the number of concepts per task can be quite high if the nature of the research problem is simple and the category is such that most of the respondents will be very familiar with it.

There is one additional point to take into account before finalizing the design. Currently, the Brand "B9" is in the market with only 2 SKUs out of the 10 to be tested. So a design was used where the new SKUs could appear or not in the board and the respondents would see anything from 2 to 10 SKUs for Brand B9. Each new SKU from Brand B9 could appear or not (Partial Profile Design).

Summarizing, we ended up with one factor with 30 levels (SKU), most of the SKUs with 2 price points (Regular and Discount), and 8 SKUs (the new ones) that should be part of a rotation group so that they do not appear all the time. Since we have labeled options (each option is a SKU), we don't need to include the factor SKU with 30 levels in the design (all the SKUs will appear all the time). What we need to do is to create factors with two levels for each SKU that will determine which level of price will be shown with each SKU. These factors will have four levels for the new SKUs, the third and fourth level meaning that the SKU will not be shown at all (rotation group). This design was generated with SAS® Proc FACTEX and Proc OPTEX, with 12 boards per respondent. Because of the rotation group each board had between 24 and 30 concepts from which one should be chosen by the respondent.

Figure 1. Screen as seen by the respondents



Brand1	Brand2	Brand3	Brand4	Brand5	Brand6	Brand7
326g Tin ○	454g Bag ○	642g Tin ○	400g Bag ○	454g Bag ○	925g Tin ○	300 Tin ○
930g Tin ○		975g Tin ○				900 Bag ○

Brand8	Brand9	Brand10	Brand11	Brand12	Brand13	Brand14
454g Bag ○	2x300g Bag ○	454g Bag ○	400 Bag ○	450 Bag ○	343 Tin ○	250g Bag ○
	2x350g Bag ○	907g Tin ○			930 Tin ○	300g Tin ○
	2x400g Bag ○					454g Bag ○
	2x450g Bag ○					
	300g Bag ○					
	326g Tin ○					
	350g Bag ○					
	400g Bag ○					
	450g Bag ○					
	915g Tin ○					

4. Analysis

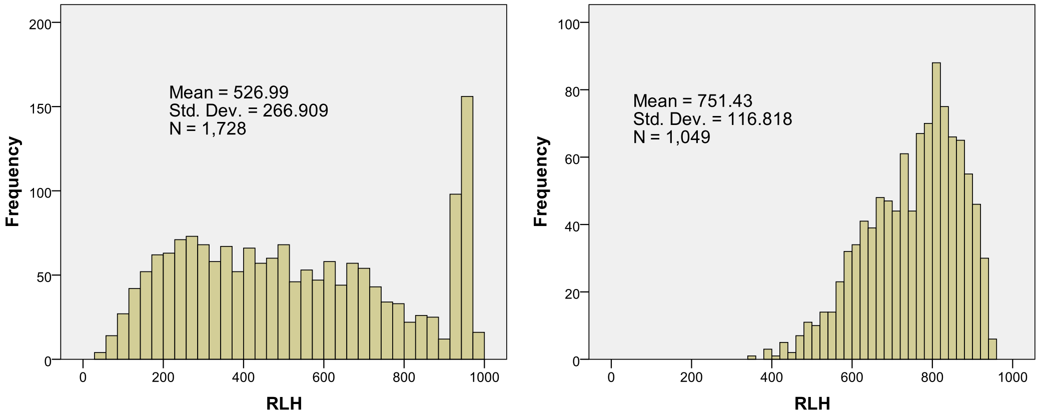
We first estimated a simple model, with two factors: SKU, estimated with effects coding, and Price, estimated with a linear function. Additionally there was the None option, which was designed as a dual response None Option.

4.1 The Hierarchical Bayes model

Let's start by looking at the Hierarchical Bayes results from Sawtooth Software's CBC/HB. The estimation process ran smoothly and converged soon, although it took us around 2 hours to run 15000 iterations on 1700 respondents. The process was likely slow because of the number of concepts per board, which when stacked into a .cho file for the Sawtooth Software CBC/HB software, becomes huge in terms of number of rows. This just means we have a lot of information from the design matrix to convey (12 boards times around 27 options per respondent). But then when one option is selected out of more than 20 in a board we end up with a dependent variable that does not provide much information for most of the concepts. This is likely the cause of the first interesting observation we found: the measure of fit RLH is much lower than we usually see in other discrete choice projects. Figure 2 compares the RLH distribution for this study with the RLH distribution from another model where we had only 5 alternatives per board. Although the average RLH is much lower here and a large proportion of respondent have RLH below 0.4 (400), the fit is pretty good relative to what we would expect by chance only – considering 25 options and no information, chance alone would get correct only around 4% of the time. This is also a reason why it is not possible to compare the RLH across studies with different number of alternatives or to judge the quality of the fit by the size of the RLH. Individual level utilities may still not be very accurate because the number of parameters per individual is much larger than the number of choices.

Although the histogram shown in Figure 2 for the traditional discrete choice design cannot be considered the standard distribution we see in a discrete choice experiment, since it varies from study to study, it does give a general idea of how one would expect the RLH to behave – a concentration toward the upper end of the range, where the fit is good. In this case, where we show many concepts, we see a quite different picture. There is a reasonable high concentration of values below 400 and a peak for values close to 1000. It seems to indicate that for many respondents we have information that is too sparse for getting a good absolute fit at individual level. RLH above 800 are usually associated with respondents who stuck to one product or brand for all the boards, regardless of the price. This was expected given the category (good purchased very often) and the fact that we were using reasonable prices ranges.

Figure 2. Distribution of the RLH fit measure for the current design (left) and a more traditional choice design (right).

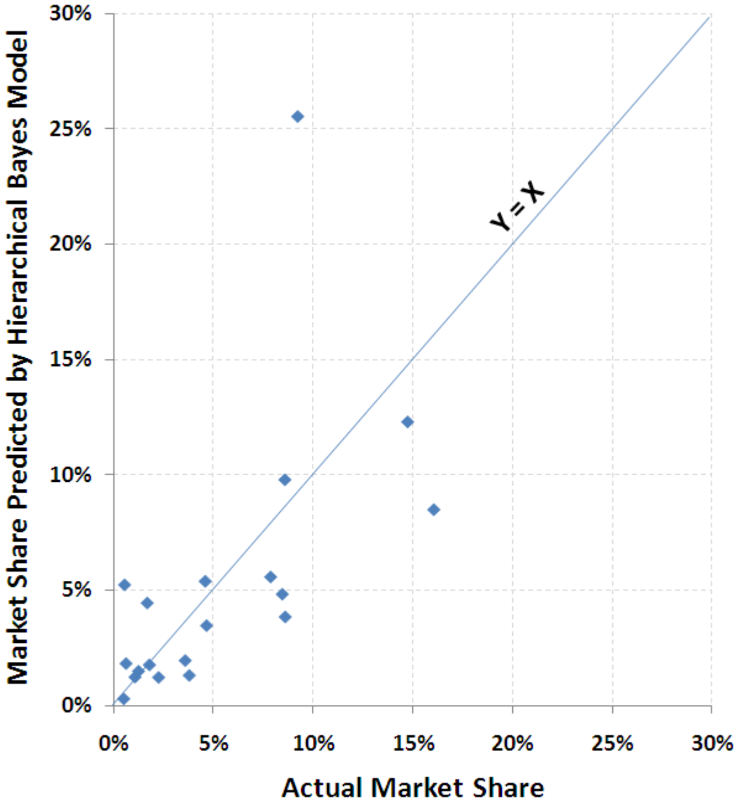


Another test of the model was done regarding the face validity of the utilities. Price was estimated with a linear function and was negative as we would expect. We also compared some products and brands and at the aggregated level the utilities were as expected.

The third assessment of the model was done through comparing Discrete Choice share against actual market shares. Although we had the actual market share for most of the SKUs, it was in Dollar value. The Discrete Choice does not give a share in dollars, so we weighted the share of preference by the frequency of purchase of each respondent to get a share of units and multiplied the units by the Dollar value of each product to get shares of Dollar value, which are shown in Figure 3. We can see that figures from Discrete Choice do not match Market Shares for most products, although we were pleased to see that the model results follow the market shares patterns in general, with the points around the line Y = X except for one outlier. Given the specific nature of the problem, different prior parameters were tested in the Sawtooth Software's CBC/HB software, all of them resulting in very similar shares, hit rates and average RLH. Because of this all the results presented are based on the default priors from CBC/HB software. We think that important factors for explaining the differences between the shares may be the fact that errors are introduced when moving from share of preference to share of Dollar value, the fact that a single choice does not account for frequency of purchase/volume purchased, the existence of store brand/generic products not accounted for in the design, the lack of precision for the market share figures (we did not get shares for all the products, for example, products 2 and 6 don't appear in Figure 3 because we did not have market share for them) and it is also possible that there is some bias in the sample, although basic demographic correction was applied.

Finally, we also looked at the percent correct classified or Hit Rate (HR) and this was the fit measure that made us feel comfortable with the model. As there was no hold-out task, one of the 12 tasks was selected randomly from each respondent to be excluded from the set of tasks used for estimation, and the model was fit with only 11 tasks per respondent. The model was then used for predicting the excluded task. The HR calculated this way was 74.8%. Considering that the model has to predict which concept will be selected in a set of 23 to 30 concepts, this HR is quite good. The high HR also may signify that differences observed matching the Market Share is not so much due to poor performance of the model. It was hypothesized that this could be caused by the over-representation of the client brand in the design. Separated models were run for the boards with different number of client concepts. Recall that the client brand had concepts that were part of the rotation group (the new concepts they wanted to test). This way the client brand was present 10 times in some boards, but only twice in others. We isolated these boards with different incidence of the client brand and fitted separated aggregated level conditional logit models to them, which always resulted in similar market shares. So the unbalance of the incidence of the client brand in the design compared to the actual market seems not to play much of a role in explaining the observed difference in the model prediction and current market share. We also noticed that over-estimated brands were often premium brands with strong presence in the market in one of more different category. As these brands are strong outside of the choice context as well as attractive for being premium, we think that respondents might have had their choice inflated regarding these brands, compared to what they would choose in the actual market. On the other hand we noticed that the model underestimated a store brand, which is a discount brand, so the same rational applies.

Figure 3. Comparing Discrete Choice share with actual Market Share. We could not get shares for some products (P2, P6 and P23) and other products don't appear in the chart because they are the new ones and as such they don't have market share (P14 to P21).

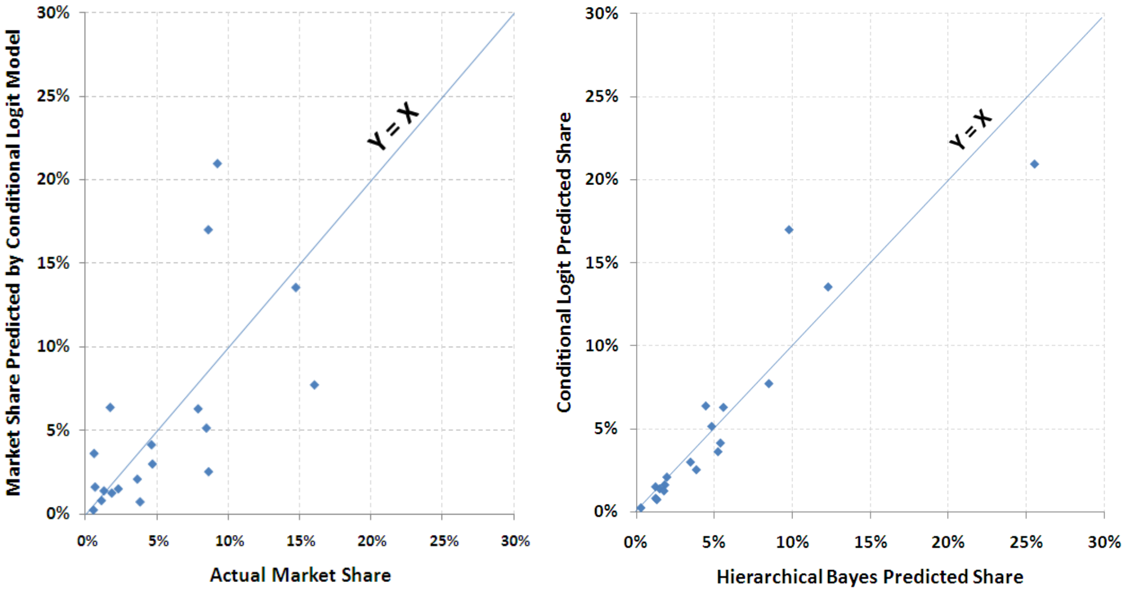


Our concern was that many concepts per board could cause poor responses since respondents would not read with attention all the concepts before selecting the most preferred one. We think this did not happen first because the category is engaging and often used, and second because there are few attributes (Brand/Price/Size/ Package Type) which made the SKU easily understood. We do not argue that having 30 concepts per board is ok for all cases, but we do think that the number of concepts per board does not need to be limited to 5 or 6 and that it depends a lot on the category and complexity of the concept as well as in the business context. We also think that showing labeled concepts, with brands fixed and representing the entire market helps make the task easier on the respondent while being also a more fair representation of the actual market. If there are 14 important brands in the market, we will pay a price in terms of lack of realism/ representativity for showing only a few brands at a time.

4.2. The Conditional Logit Model

Although the Hierarchical Bayes model (and especially its Sawtooth Software implementation) has been an important advancement in Discrete Choice modeling, changing the way discrete choice is seen and used in Marketing Research, it is still one of the most complex choice models. The concepts of priors, Hierarchical Bayes modeling and MCMC simulations, interpretations of parameters, effects, interactions and fit statistics are not simple nor easily understood. These considerations along with the fact that the RLH was not high for many respondents made us wonder about how the Conditional Logit Model (CLM) would fit this data, since it is an aggregated level model. The CLM is also appealing because we don't really need respondent level utilities in this specific case.

Figure 4. Comparison between Conditional Logit Model and Hierarchical Bayes as well as between Conditional Logit and Hierarchical Bayes model.



The CLM was fit using the package "mlogit" from R software. If we use the Mean Squared Error (MSE) to compare CLM and HB in terms of predicting Market Share, the CLM will win (MSE = 0.00198 against MSE=0.00218). But Figure 4 (right side) shows that there is hardly any meaningful difference between market shares predicted by the Conditional Logit Model and Hierarchical Bayes. This shows that, at least when we look at Market Shares, the CLM is robust to the violation of the assumption of independence of responses and the within respondent effect. At the individual respondent level the CLM performs very poorly compared with HB with Hit Rate of only 28.5% (against 74.8% for HB). This was expected since the CLM imposes the same model on every respondent.

The CLM can also be useful in helping to improve the model through different model specification, because it is easy to test coefficients with the CLM. For example, each of the 30 products has a size in grams attached to its description. It makes sense to think that there could be an interaction between price and size (let's call it Price*Size) because the perception of price should be affected by the size of what we buy. The CLM allow us to test whether or not the interaction Price*Size improves significantly the model. Table 2 shows the Log Likelihood for the model with price only and how it increases when we include interactions in the model. The statistic -2ln(LR), where LR = Likelihood Ratio, has a Chi-Square distribution with 1 degree of freedom (under some assumptions) and as such can be tested. We can see that the inclusion of the interaction Price*Size leads to a significant increase in the Log Likelihood.

Table 2. Improvement in the model by including interactions

	Log Likelihood	-2ln(M0/M1) ~$X^2(1)$	$P(X^2(1)>X)$
Price Only (M0)	-50255.13		
Price and Price* Size (M1)	-50251.8	6.66	0.010
Price and Price/Size (M1)	-50251.63	7.00	0.008

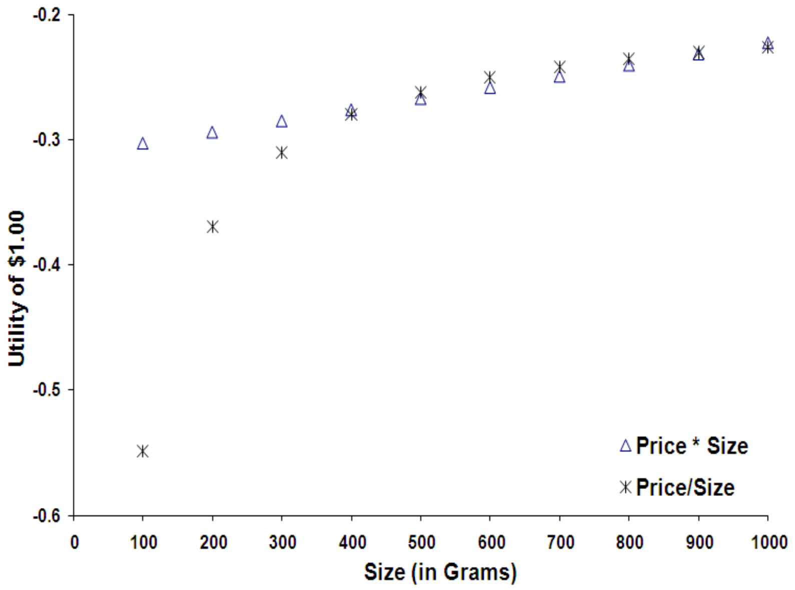
The Price*Size interaction is not very easy to interpret because it has no clear meaning. But another interaction between Price and Size that is meaningful can be created if we divide price by size instead of multiplying (let's call it Price/Size). In this case the interaction term is the price per unit of size which is something many of us think about when comparing prices between same products with different sizes. We tested the Price/Size term and it too ended up being significant with an increase in the log Likelihood that is a little larger than when using the multiplicative interaction, as can be seen in Table 2. The CLM also showed us, through the test for the interaction coefficient, that both the multiplicative and ratio terms are significant in the model, which is an equivalent test to the one in Table 2.

It is easier to understand the effect of these interaction terms and grasp the difference between the multiplicative and ratio interactions by plotting them. Figure 4 shows the effect of $1.00 increase in price at some levels of the size attribute. Remember that the presence of the interaction means that it is meaningless to interpret the coefficient of Price alone, because the effect of price depends on size, so this is what we are looking at in Figure 5. The Price*Size interaction means that the utility of price drops with an increase of $1.00 (utility is always negative in Figure 5) but the drop decreases when the size increases. Notice that this makes sense: increasing the price by $1.00 is bad if the size is small, but not so bad if the size is large. In other words, it is worse increasing the price by a certain amount if the size is small. Both the multiplicative and the ratio terms lead to this same interpretation, but the ratio term assumes a non-constant variation.

Additionally the CLM also allows for direct interpretation of the exponentiated coefficients as change in the Odds of selecting a product. For example, according to the ratio interaction the utility of increasing Price by $1.00 in a 200g product is -0.37. That is, increasing the Price by $1.00 makes the Odds of selecting the product decrease 0.69 times or around 30% (Exp(-0.37) = 0.69).

The fact that the Price/Size term is directly interpretable makes it meaningful to have it in the model even without the price main effect term. However, excluding price leads to a significant drop in the Log Likelihood (-50301.82). So we decided for including Price and Price/Size in our final model. We ran the HB Model again, this time including Price/Size and we noticed a marginal increase in the Hit Rate from 74.8% to 75.6%.

Figure 5. How the utility of increasing the Price by $1.00 changes according to the Size and the type of interaction



4.3. The Effect of Number of Boards

In order to develop some understanding of how many boards we should use, we decided to re-estimate the model with fewer tasks and see how the Hit Rate would turn out. In this case the ideal analysis is to consider the initial boards for estimation because this way we can answer the question "How would the model have fit had we asked only [say] 10 instead of 12 boards?" Answering this question will allow us to understand if when we have many options the number of boards should be increased or not. This means we cannot select randomly boards to be excluded from the estimation set and predicted afterwards because if we do that many excluded boards will be among the [say] 10 first boards. So what we did here is that we always tried to predict board 12 which was never in the estimation set.

Figure 6 shows the Hit Rate with and without None for models estimated from the first 5 boards to the first 11 boards. Excluding the None option is interesting because it has a high utility and around 35% of the Share of Preference. So it is not very difficult to predict it correctly and we could get 35% correct just by predicting that everybody chooses None. We can see that although the Hit Rate decreases, it is still high even if we use only 5 boards. Figure 7 adds to this conclusion showing that the coefficient estimates are quite similar whether we use 5 or 12 boards. We also noticed (but we are not showing here) that the standard error of the estimates, as calculated by the standard deviation of the respondent level shares divided by the square root of the sample size, are very similar (they don't seem to depend on the number of boards).

We conclude from this that there is not much information gained from boards 6 to 12. We think this could mean that there are essentially two types of respondents: those who are loyal to the product and will usually select the same product and this will be already clear from the 5 initial tasks so that for this type of respondent the model will be good even with only 5 tasks. The other type would be the non-loyal respondents, for whom the model would not fit well regardless of the number of boards. The other factor that could play a role here is that we have only two attributes (Product and Price) and this may make it easier to determine the preference of respondents from the initial set of boards. We think this conclusion could be very specific to our study and would not recommend decreasing the number of boards until more research is done on the subject.

Figure 6. How good would be the HB Model in terms of HR had we used fewer boards.

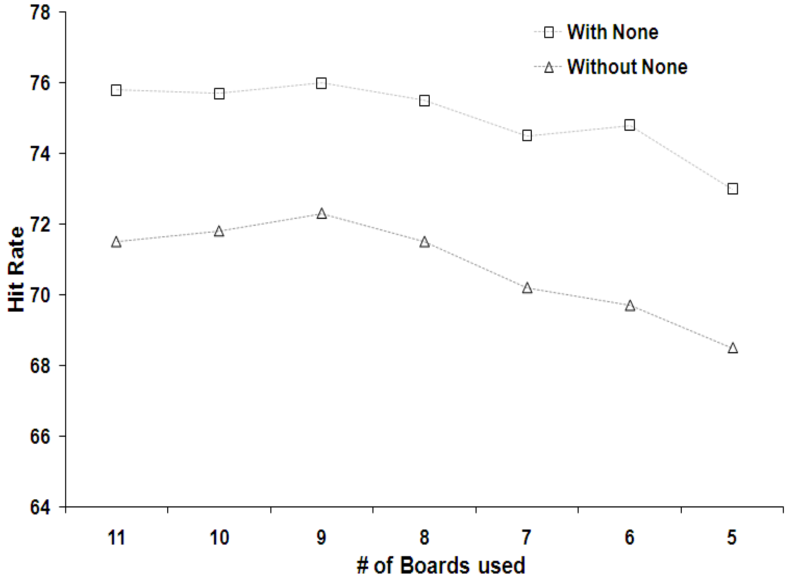
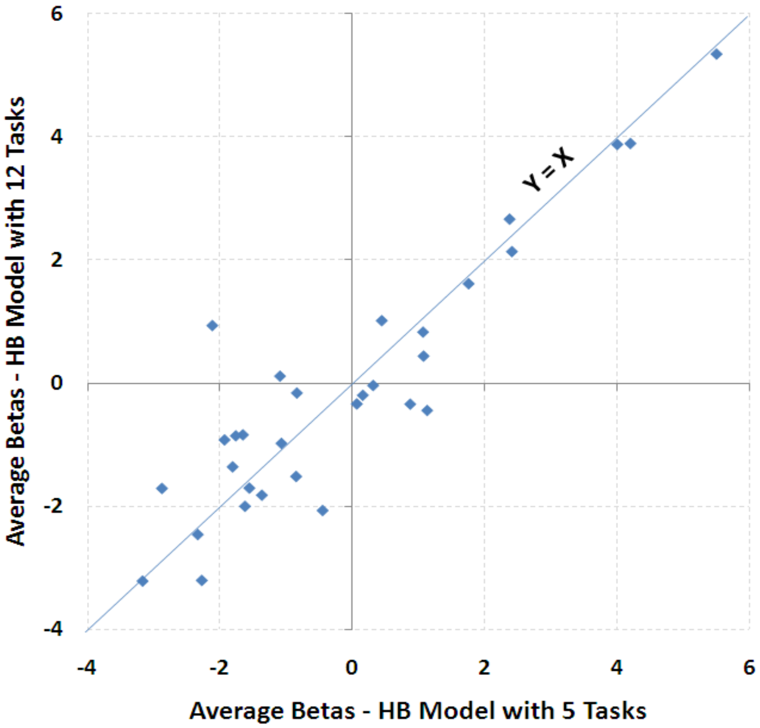


Figure 7. Betas estimates with 5 and 12 boards.



5. Conclusions

There are not many studies exploring the effect of the number of options per board. While we too don't go as far as testing different number of options per board to come up with recommendations as to what the ideal number is, we do explore an extreme case where between 23 and 30 options are shown at one time. The results are interesting and provide some guidelines.

Our case study is from the food industry and is about products that consumers buy often. That means that this sort of decision is one that they are familiar with. There are only 4 attributes which also helps to make the concept easy to understand. Although the number of options is way above what is usually recommended, we noticed that respondents can still provide good answers since the decision is still similar to the one they do regularly.

The individual level RLH was lower than in studies with fewer options as we should expect and the Hit Rate was quite good. We think that the low RLH, although not reflective of bad fit, can indicate the lack of individual level information on some levels of some attributes therefore making the use of individual level results subject to caution. The predictions of Market Share were also reasonable. The simple Conditional Logit model worked well and we think that in this kind of situation, if we don't feel comfortable using Hierarchical Bayes at individual Level, it is ok to work with aggregated level models like the Conditional Logit. The aggregated level logit model can help specifying the model with interactions, for example, although external theory as to whether or not the specification is meaningful should always be considered.

The number of options to be shown per board should depend on the complexity of the concepts. We think there is always a value on showing more options if by doing so we can represent the market accurately and especially if the alternatives are labeled (usually labeled alternatives happen when we have as many brands as options so that each option refers to a brand and all brands appear all the time). Labeled options make the choice task easier since the brands are always the same across tasks.

The final comment is about the number of boards. We had 12 boards but our studies showed that we did not need more than 5 or 6 boards. It seems that if the product is something with which respondents are very familiar and can easily make choices, then even with so many options we will not need many boards because the respondent's preference will be clear from the initial boards so that the later ones will not add much information. We also speculated that the lack of improvement of the model when we go past 5 or 6 boards could be due to the fact that the decision rule does not follow the logit model, like some sort of non compensatory decision. Yet the fact that the Hit Rate was good with only a few boards points to the need of more study in this area because it could also be that overall there is no need for more than a few boards when doing a Conjoint study in familiar categories.

REFERENCES

Cunningham, C. E., Deal, K., Neville, A. & Miller, H. (2004): Modeling Conceptually Complex Services: The Application of Discrete Choice Conjoint, Latent Class, and Hierarchical Bayes Analysis to Medical School Curriculum Redesign - *Proceedings of the Sawtooth Software Conference, pp. 93 – 106*.

Pinnell, J. & Englert, S. (1997): The Number of Choice Alternatives in Discrete Choice Modeling - *Proceedings of the Sawtooth Software Conference, pp. 121 – 153*.

Zwerina, K., Huber, J. & Kuhfeld, W. (1996): A General Method for Constructing Efficient Choice Design – *SAS Technical Support Documents.*

LEVERAGING THE UPPER LEVEL MODELS IN HB FOR INTEGRATED STAKEHOLDER MODELING

BRIAN GRINER[35]
QUINTILES MARKET INTELLIGENCE,
QUINTILES CONSULTING
IAN MCKINNON[36]
KANTAR HEALTH
PIETER SHETH-VOSS[37]
PROVEN, INC.
JEFFREY NIEMIRA[38]
QUINTILES MARKET INTELLIGENCE,
QUINTILES CONSULTING

LINKED MODELING

Linked modeling involves several steps. The first step in building an integrated demand model is to identify the decisions made by each stakeholder (dependent variables). The second step is to determine the likely degree of influence on product adoption each stakeholder will have (should they be in the model?).

This application focuses on the decisions of three key stakeholders:

1. **Payers**
   – What is the level and type of coverage/reimbursement a new product should have if any?
2. **Physicians**
   – Should I prescribe/recommend a new product and if so to what patients?
3. **Patients**
   – I just saw an ad for a new product. Should I ask my doctor if it is appropriate for me?

   – If a doctor recommends a new product at a higher cost, should I accept it or ask for an older less expensive drug?

STEP 1: IDENTIFYING STAKEHOLDER DECISIONS

In the conceptual model of the pharmaceutical marketplace (See Figure 1), each stakeholder has his/her own set of concerns and motivations, while being influenced by, and influencing the other stakeholders:

> **Government and regulatory agencies**, like the FDA, approve new manufacturers' therapies, place restrictions on manufacturers – what's in the label, promotional

[35] Chief Methodologist, Quintiles Market Intelligence, Quintiles Consulting
[36] Chief Research Officer, Kantar Health
[37] President, Proven, Inc.
[38] Senior Analyst, Quintiles Market Intelligence, Quintiles Consulting

messages – NOT off-label; enforce coverage mandates for managed care; and provide protection for patients that access services.

Manufacturers negotiate contracts for new therapies with managed care (and government), and promote new therapies to physicians and patients.

Managed Care providers create and manage health care plans for employers.

Patients & caregivers seek the advice of physicians on appropriate therapy options based on efficacy, safety, tolerability, etc; react to the cost of therapy options; and respond to product promotions by asking physicians about specific therapy options.

Physicians evaluate new therapy options based on product features: efficacy, safety, tolerability, mechanism of action, dosing, formulation, cost to patient and practice, etc.; they react to patients' questions and requests for new therapy options; and they assess the impact of product reimbursement on their overall business risk (e.g., stocking vs. write-and-refer).

Payers evaluate new therapies for inclusion on their formulary based on efficacy, safety, tolerability, etc., and determine the level of coverage / reimbursement for new therapies based on unmet need in the category, degree of innovation and (increasing) comparative effectiveness.

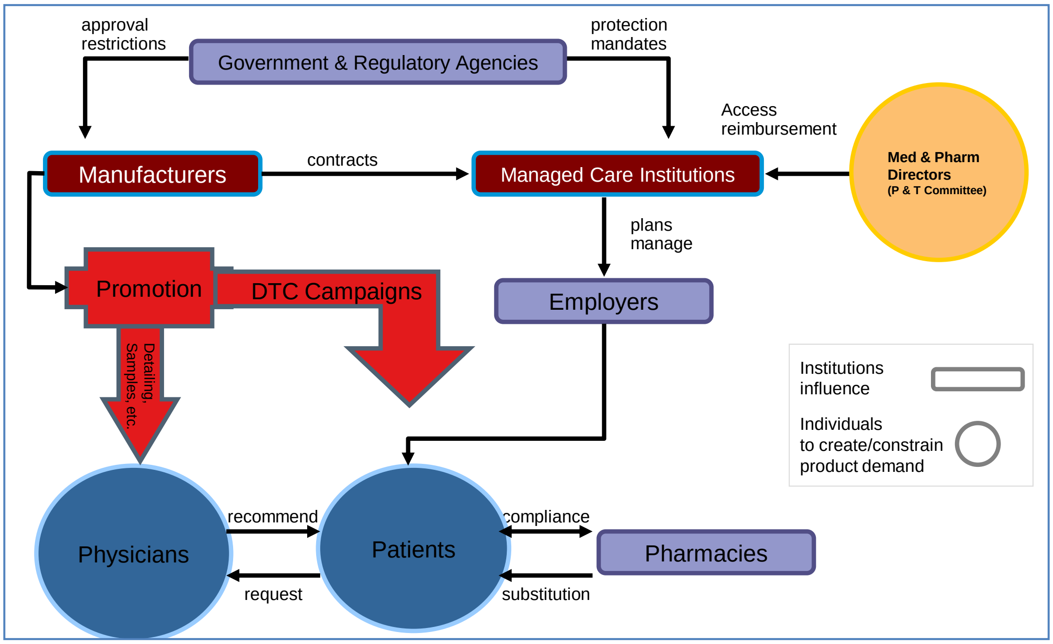
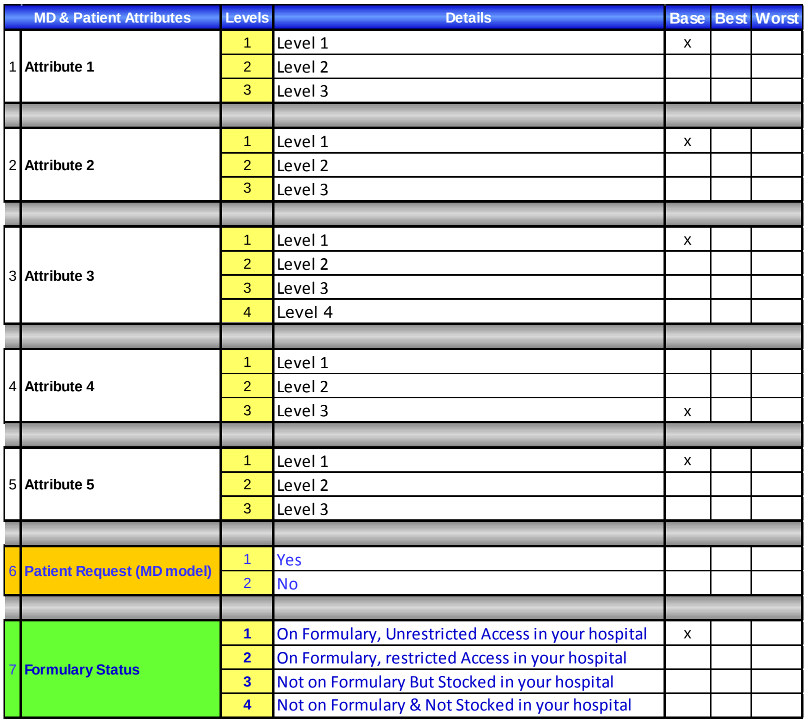


Figure 1: Example of Conceptual Model: US Pharmaceutical Marketplace

Step 2: Experimental Design

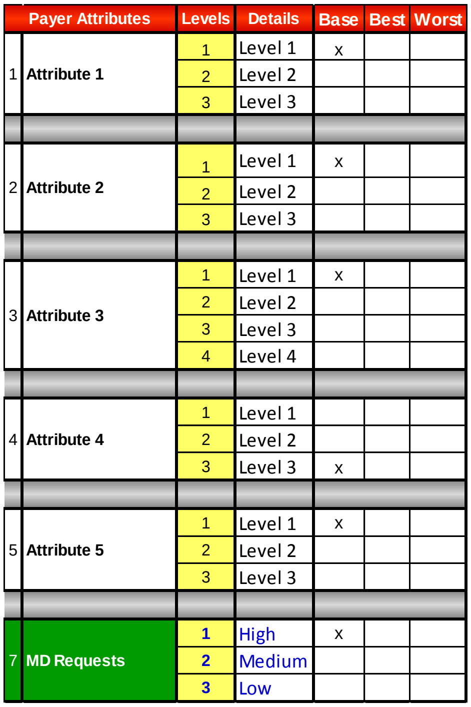
Experimental designs are different for each group. The physician and patient designs both include profile attributes and payer DV (e.g., formulary status. See Figure 2). The physician design also includes patient DV (e.g., patient requests).



	MD & Patient Attributes	Levels	Details	Base	Best	Worst
1	Attribute 1	1	Level 1	x		
		2	Level 2			
		3	Level 3			
2	Attribute 2	1	Level 1	x		
		2	Level 2			
		3	Level 3			
3	Attribute 3	1	Level 1	x		
		2	Level 2			
		3	Level 3			
		4	Level 4			
4	Attribute 4	1	Level 1			
		2	Level 2			
		3	Level 3	x		
5	Attribute 5	1	Level 1	x		
		2	Level 2			
		3	Level 3			
6	Patient Request (MD model)	1	Yes			
		2	No			
7	Formulary Status	1	On Formulary, Unrestricted Access in your hospital	x		
		2	On Formulary, restricted Access in your hospital			
		3	Not on Formulary But Stocked in your hospital			
		4	Not on Formulary & Not Stocked in your hospital			

Figure 2: Physician and Patient Linked Model Design

The payer design usually includes the physician DV (e.g., physician requests); however, some payer designs do not include a link directly from physicians (See Figure 3.).



Payer Attributes		Levels	Details	Base	Best	Worst
1	Attribute 1	1	Level 1	x		
		2	Level 2			
		3	Level 3			
2	Attribute 2	1	Level 1	x		
		2	Level 2			
		3	Level 3			
3	Attribute 3	1	Level 1	x		
		2	Level 2			
		3	Level 3			
		4	Level 4			
4	Attribute 4	1	Level 1			
		2	Level 2			
		3	Level 3	x		
5	Attribute 5	1	Level 1	x		
		2	Level 2			
		3	Level 3			
7	MD Requests	1	High	x		
		2	Medium			
		3	Low			

Figure 3: Payer Model

Traditional Linked Model Approaches: Non-Recursive vs. Recursive Models

Recursive Linked Model

The recursive model excludes the link from physicians to payers. Instead market demand is estimated by using the ***predicted outcomes of payers and patients*** to calculate expected ***utilities*** in the ***physician model*** (See Figure 4).

Payer model = Pr(Tier/medical reimbursement of Product X | E(MD demand), attribute levels, availability of new competitive products)

Physician model = Pr(Prescribe Product X | attribute levels, E(tier/medical reimbursement), E(patient requests), availability of new competitive products)

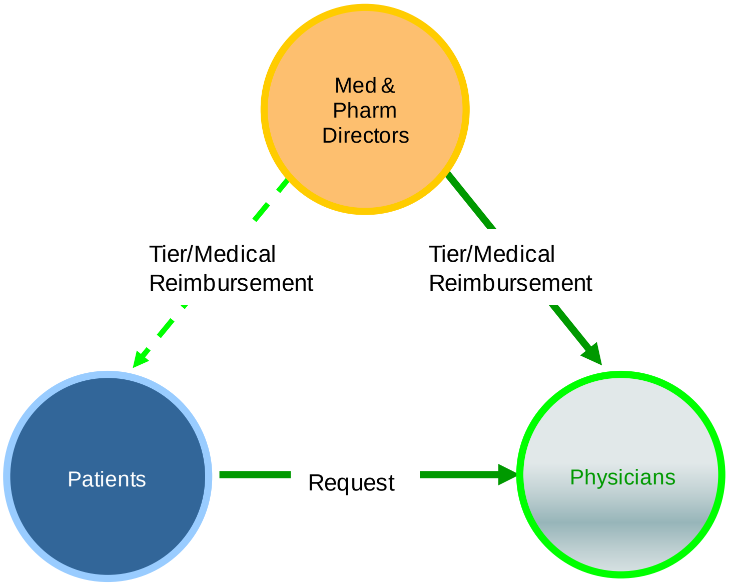
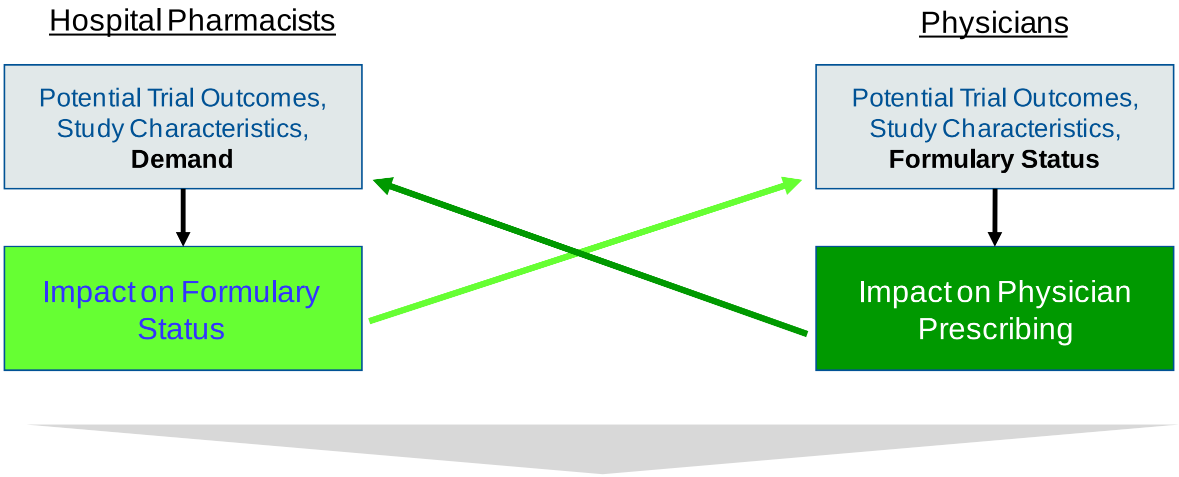
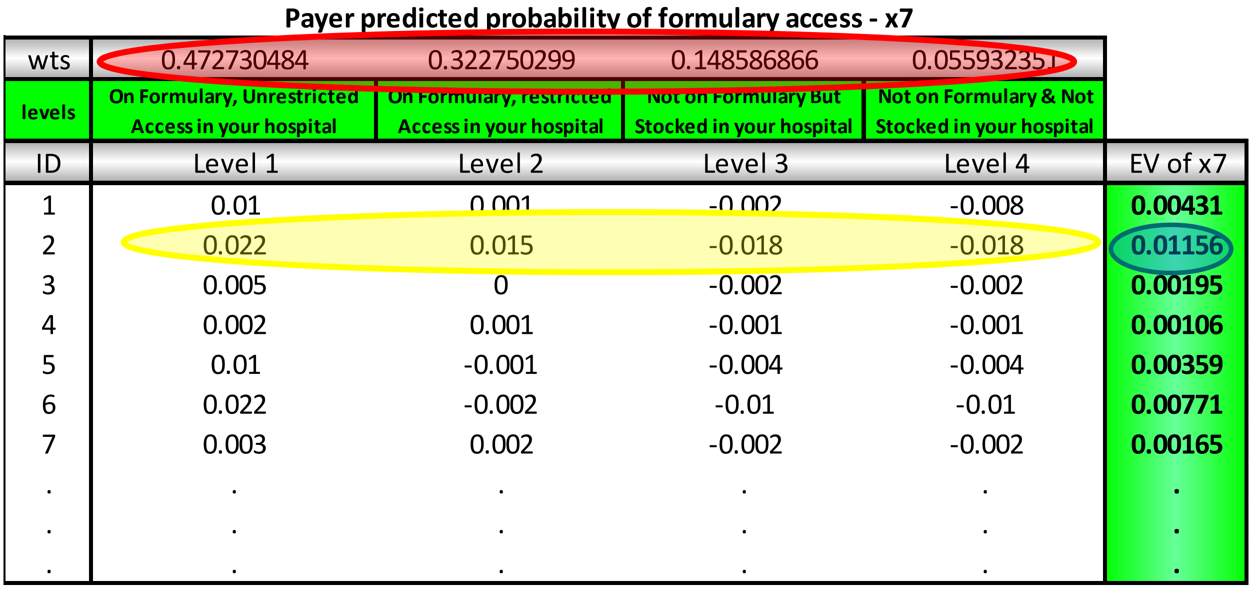


Figure 4: Recursive Linked Model

Non-Recursive Linked Model

The non-recursive model includes feedback loops from physicians to payers and from payers to physicians in the simulator in order to calculate the expected model utilities given the stakeholder's response to the product profile (See Figure 5). The non-recursive model must iterate until it reaches an equilibrium level of demand and formulary status.





Payer predicted probability of formulary access - x7

wts	0.472730484	0.322750299	0.148586866	0.055932351	
levels	On Formulary, Unrestricted Access in your hospital	On Formulary, restricted Access in your hospital	Not on Formulary But Stocked in your hospital	Not on Formulary & Not Stocked in your hospital	
ID	Level 1	Level 2	Level 3	Level 4	EV of x7
1	0.01	0.001	-0.002	-0.008	**0.00431**
2	0.022	0.015	-0.018	-0.018	**0.01156**
3	0.005	0	-0.002	-0.002	**0.00195**
4	0.002	0.001	-0.001	-0.001	**0.00106**
5	0.01	-0.001	-0.004	-0.004	**0.00359**
6	0.022	-0.002	-0.01	-0.01	**0.00771**
7	0.003	0.002	-0.002	-0.002	**0.00165**
.	.	.	.	.	**.**
.	.	.	.	.	**.**
.	.	.	.	.	**.**

Figure 5: Non-Recursive Linked Model example – payer link to physician model

In Figure 5, the expected utility of formulary status in the physician model are calculated using the predicted probabilities from the payer model as weights. A similar calculation is done in the payer model to calculate the expected utility of the level of patient demand using the predicted shares as weights. To ensure that the non-recursive model obtains a unique solution, the conditions for model identification must be satisfied.[39] In this example, the conditions for identification are satisfied by the inclusion of the dependent variable from the payer model as linking attributes in the physician design and vice versa (patient model is recursive and is identified). Models with more complex linkages should be checked for identification to ensure a unique equilibrium solution is possible. The equations in the non-recursive model are given below.

Payer model = Pr(Tier/medical reimbursement of Product X | attribute levels, availability of new competitive products)

- Direct effect on *patients* (requests) and *physicians* (prescribing)
- Indirect effect on *physicians* through patients (requests | tier/med reimbursement)

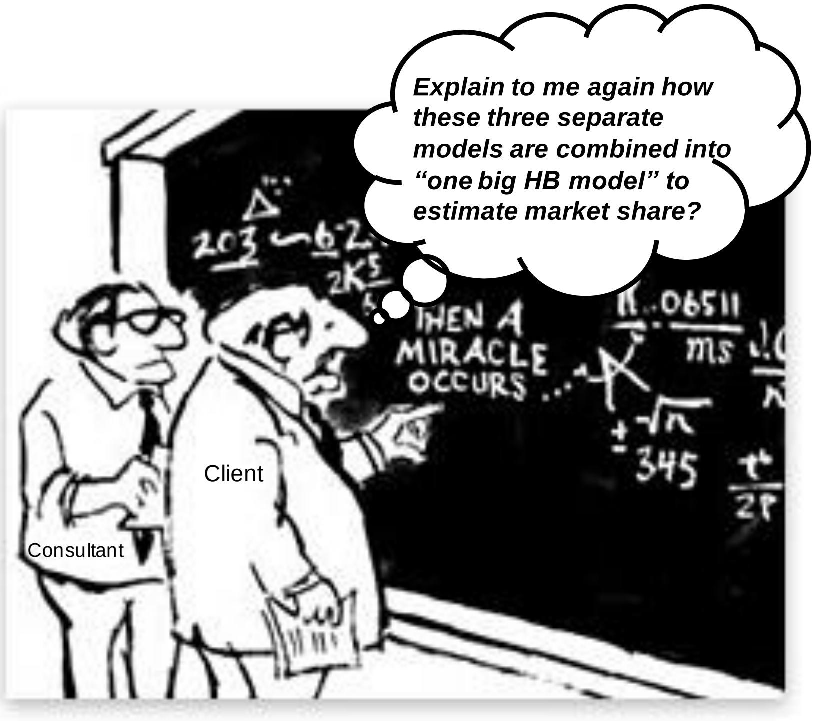
Patient model = Pr(Request Product X | attribute levels, E(tier/medical reimbursement), availability of new competitive products)

- Direct effect on *physicians* (prescribing)

Physician model = Pr(Prescribe Product X | attribute levels, E(tier/medical reimbursement), E(patient requests), availability of new competitive products)

[39] Judge, Hill, Griffiths, Lutkepohl and Lee, 1988, "Introduction to the Theory and Practice of Econometrics," see chapter 14.6 "Identifying an Equation within a System of Equations pp.622-626 and chapter 14.7 "Some Examples of Model Formulation, Identification and Estimation" pp. 626-630 for the rank and order conditions for identification of a system of equations.

UNDERSTANDING THE HB LINKED STAKEHOLDER MODEL


Explain to me again how these three separate models are combined into "one big HB model" to estimate market share?
THEN A MIRACLE OCCURS...
Client
Consultant

THE BASIC HIERARCHICAL BAYESIAN REGRESSION MODEL

The basic HB model starts with a *"random effects"* framework. The utility for respondent *i* and attribute *k* follow a distribution of utilities in the study population (e.g., a normal distribution) with a mean utility mu_k and variance $sigma^2_k$.

$$\beta_{ik} \sim N(\mu_k, \sigma_k^2)$$

THE UPPER-LEVEL MODEL

We can extend the basic HB model by conditioning the mean vector of *k* utilities as a function of respondent-level covariates (e.g., specialty, country, etc.) using a multivariate regression.

$$\beta_i = \Theta' z_j + \varepsilon_j$$

- Where theta is a vector of mean utilities for *k* attributes for subject *i,*
- Theta is a *k x j* matrix of coefficients conditioning *k* betas on *j* covariates
- *Z* is the vector of *j* covariates for subject *i*
- Epsilon is the vector of *j* random errors for subject *i* assumed to be distributed normal ~ N(0, Δ)

The Upper-Level Model allows the respondent level utilities to *"shrink"* to the mean of their peers as defined by the covariates.

THE HIERARCHICAL BAYES LINKED STAKEHOLDER REGRESSION MODEL

The Linked HB model looks a lot like the basic model with covariates in the Upper-Level Model. The differences are:

- **How we define the dependent variable for each stakeholder group.** For example,
  - Physicians make therapy choices for patients
  - Payers make market access decisions
  - Patients make decisions to request therapies or accept therapy recommendations from physicians
- **How we define the term *"covariates."*** For example,
  - Instead of modeling specialty within a stakeholder group, we model the stakeholder groups directly (e.g., physicians, payers, patients)
- **How we specify the means in the Upper-Level Model (i.e., theta in equation 2)**
  - Use of interaction terms (or super attributes) to shrink to the appropriate means in the ULM , For example, to model "country" level differences within stakeholder groups, we need to include interaction terms between stakeholder groups and countries (e.g., ***beta*** is a function of (physician / payer / patient)*country)

In the Linked HB model, the Upper-Level Model now links different stakeholder groups:

- ***Theta*** models differences in stakeholder utility for specific attributes directly
- ***Delta*** models stakeholder preferences indirectly by modeling relationships between attribute utilities that are conditioned on membership to different stakeholder groups

For example, ***Theta*** may show that physicians, patients and payers all have positive utilities for efficacy, and physicians and patients have higher utility for IV versus IM injection, while payers show the opposite. ***Delta*** may show a positive relationship between efficacy and IV indicating the physician and patient preferences for IV offset the payer preferences for the (cheaper) IM injection. The beauty of using the Upper-Level Model in the HB framework for the Linked Stakeholder Model is that the coefficients of this model can be estimated in one carefully constructed HB model using currently available software (i.e., CBC HB). The choice data for each stakeholder is stacked in one dataset. The linking attributes that are not present in every design are set to zero if using effects coding or missing in other coding schemes. The dependent variables are recoded or rescaled depending on the type of model being estimated (e.g., formulary status is recoded to a Yes/No in a discrete choice model or 0/100 in an allocation model). The demographic data is coded so that the means of the betas shrink to the correct stakeholder group in the Upper-Level Model (e.g., if the study contains two patient types and two physician specialties plus payers then a super attribute can be constructed for the Upper-Level Model with levels: 1 - patient type 1/specialty 1; 2 - patient type 1/specialty 2; 3 - patient type 2/specialty 1; 4 - patient type 2/specialty 2; 5 – payers). Unique ids need to be created for each stakeholder group (and patient type if multiple patient types are evaluated) so that the modeling software correctly identifies the appropriate unit of analysis for the choice model.

HIERARCHICAL BAYES LINKED STAKEHOLDER MODEL CONSIDERATIONS

Linked Hierarchical Bayesian stakeholder demand modeling has pros and cons.

PROS

- Greater insight into **drivers and barriers** to product demand with **specific stakeholders groups**
- **Upper-Level Model** allows for greater insight into **similarities and differences** in utilities among different stakeholder groups
- The research design for payers can **accommodate a diverse set of market access and reimbursement factors** such as copay level, prior authorizations, step edits, etc. to identify key drivers and optimal strategies for formulary acceptance/placement that can be tested directly with physicians and patients
- Provides **greater statistical support** for estimation of models for small stakeholder groups (i.e., payers)

CONS

- Increased sample requirements and **cost**
- **Payers can be difficult to recruit in US** leading to smaller sample sizes
- Payers are even more difficult to recruit outside the US **therefore different modeling strategies are required to incorporate ex-US payers** in the integrated demand model

In closing, to demonstrate the value of these tradeoffs, consider the following case study.

CASE STUDY: FORECASTING DEMAND WITH NOVEL PRODUCT IN MATURE BUT COMPETITIVE MARKET

MARKET CONTEXT

A major pharmaceutical manufacturer is planning for the launch of a new product that represents a new formulation and different mechanism of action in a very mature market, when several new products will be entering the market in the near future. Additionally, their product's safety profile is such that it will require close monitoring of the patient which has implications for managed care (higher overall costs).

BUSINESS GOAL

In addition to wanting to forecast the overall demand for new product launch to inform strategic planning, they also want to:

Measure adoption in different patient populations and stakeholder groups (payers, patients, physicians)

Understand the impact of competitors on product demand and adoption rates

Measure the impact of monitoring requirements on managed care adoption and coverage levels

Test the impact of patient assistance programs on adoption with different stakeholders

HB Linked Model Considerations

In order to link these different stakeholder groups and answer these business questions, Quintiles designed HB Linked models for each stakeholder that estimated the differences in stakeholder utility for specific attributes directly, and modeled relationships between attribute utilities to model stakeholder preference.

The model required:

2 linking attributes (coverage level and patient request),

2 managed care attributes (step edit, and distribution channel), and

3 product attributes (efficacy, adverse event, and safety requirement)

Specific considerations for stakeholder attribute sensitivities were as follows:

Physician Model Attribute Sensitivities

- **Managed care attributes are important to physicians**
  - Coverage level is ranked 3rd among all the attributes
  - Distribution channel is ranked 5th
  - MC Restrictions (Step Edit) is ranked 9th in importance
- **Safety requirements are important to physicians but not as much as coverage level**
  - Safety 1 (patient monitoring) is ranked 7th in importance
- **Patient requests for new drugs are NOT important to physicians**
  - Patient request is ranked 18th (of 23 attributes) in importance to physician prescribing
  - Coverage level is 5 times more important than patient request in predicting the impact of physician prescribing
- **Efficacy is more important to physicians than Adverse Events**
  - 3 of the 4 efficacy attributes are ranked higher than all but 4 of the Adverse Events

Patient Model Attribute Sensitivities

- **Managed care attributes are important to patients, too!**
  - Coverage level is ranked 1st among all the attributes
  - Distribution channel is ranked 4th
  - MC Restrictions (Step Edit) is ranked 9th in importance

- **Safety requirements are NOT as important to patients as they are to physicians**
    - Safety 1 (patient monitoring) is ranked 13th in importance
    - Implication: Monitoring requirements are considered a cost of entry in order to receive the benefits of the new therapy
- **Efficacy is important to patients but so are Adverse Events!**
    - 2 of the 4 efficacy attributes are ranked in the top 10 but 5 of the top 10 attributes are Adverse Events

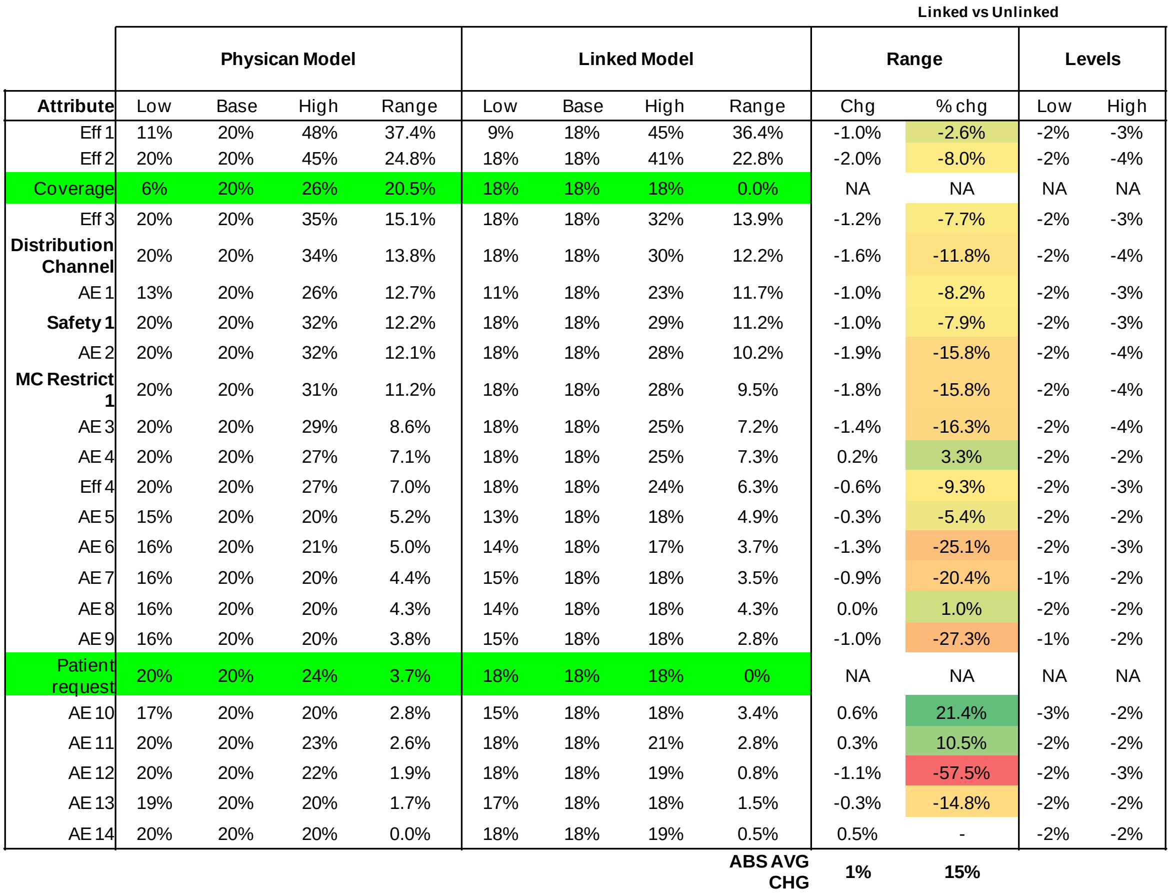
PAYER MODEL ATTRIBUTE SENSITIVITIES

- **Adverse events are the MOST important attributes to payers!**
    - Payers care the most about the COST of therapy
    - What is important to physicians and patients is NOT what is important to payers
        - The 2nd most important attribute to physicians (Eff 2) is ranked LAST by payers
        - The MOST important attribute to payers (AE 8) is ranked 16th by physicians
- **Payers seem to have more in common with patients than physicians???**
    - The MOST important attribute to payers (AE 8) is ranked 8th by patients (vs 16th by physicians)
    - The 1st and 2nd most important attributes to patients (Eff1 and Eff 3) are also in the top 10 payer attribute ranking (5th and 8th respectively)
- **Why are payers more aligned with patients than physicians?**
    - Payers care the most about the COST of therapy
    - Payers (and most employers) tend to shift the cost of therapy to the patient (employee)
    - What matters MOST to patients? Coverage level (i.e., COST)

LINKED MODEL ATTRIBUTE SENSITIVITIES

The linked model is a modified version of the physician model.

- The linking attributes are used to integrate the physician utilities
    - That’s why the ranges are 0%
- The linked model will usually shrink the sensitivity of the physician model – but NOT always
    - The largest changes tend to occur where the largest disparities exist between payers and physicians
    - AE 10, 11, 4 and 8 actually INCREASE in sensitivity somewhat



									Linked vs Unlinked			
	Physican Model				Linked Model				Range		Levels	
Attribute	Low	Base	High	Range	Low	Base	High	Range	Chg	% chg	Low	High
Eff 1	11%	20%	48%	37.4%	9%	18%	45%	36.4%	-1.0%	-2.6%	-2%	-3%
Eff 2	20%	20%	45%	24.8%	18%	18%	41%	22.8%	-2.0%	-8.0%	-2%	-4%
Coverage	6%	20%	26%	20.5%	18%	18%	18%	0.0%	NA	NA	NA	NA
Eff 3	20%	20%	35%	15.1%	18%	18%	32%	13.9%	-1.2%	-7.7%	-2%	-3%
Distribution Channel	20%	20%	34%	13.8%	18%	18%	30%	12.2%	-1.6%	-11.8%	-2%	-4%
AE 1	13%	20%	26%	12.7%	11%	18%	23%	11.7%	-1.0%	-8.2%	-2%	-3%
Safety 1	20%	20%	32%	12.2%	18%	18%	29%	11.2%	-1.0%	-7.9%	-2%	-3%
AE 2	20%	20%	32%	12.1%	18%	18%	28%	10.2%	-1.9%	-15.8%	-2%	-4%
MC Restrict 1	20%	20%	31%	11.2%	18%	18%	28%	9.5%	-1.8%	-15.8%	-2%	-4%
AE 3	20%	20%	29%	8.6%	18%	18%	25%	7.2%	-1.4%	-16.3%	-2%	-4%
AE 4	20%	20%	27%	7.1%	18%	18%	25%	7.3%	0.2%	3.3%	-2%	-2%
Eff 4	20%	20%	27%	7.0%	18%	18%	24%	6.3%	-0.6%	-9.3%	-2%	-3%
AE 5	15%	20%	20%	5.2%	13%	18%	18%	4.9%	-0.3%	-5.4%	-2%	-2%
AE 6	16%	20%	21%	5.0%	14%	18%	17%	3.7%	-1.3%	-25.1%	-2%	-3%
AE 7	16%	20%	20%	4.4%	15%	18%	18%	3.5%	-0.9%	-20.4%	-1%	-2%
AE 8	16%	20%	20%	4.3%	14%	18%	18%	4.3%	0.0%	1.0%	-2%	-2%
AE 9	16%	20%	20%	3.8%	15%	18%	18%	2.8%	-1.0%	-27.3%	-1%	-2%
Patient request	20%	20%	24%	3.7%	18%	18%	18%	0%	NA	NA	NA	NA
AE 10	17%	20%	20%	2.8%	15%	18%	18%	3.4%	0.6%	21.4%	-3%	-2%
AE 11	20%	20%	23%	2.6%	18%	18%	21%	2.8%	0.3%	10.5%	-2%	-2%
AE 12	20%	20%	22%	1.9%	18%	18%	19%	0.8%	-1.1%	-57.5%	-2%	-3%
AE 13	19%	20%	20%	1.7%	17%	18%	18%	1.5%	-0.3%	-14.8%	-2%	-2%
AE 14	20%	20%	20%	0.0%	18%	18%	19%	0.5%	0.5%	-	-2%	-2%
								ABS AVG CHG	**1%**	**15%**		

Figure 6: Comparison of Linked and Unlinked Models

STUDY TAKEAWAYS

Big disconnects between physician drivers of prescribing and payer drivers of access (i.e., coverage level – see Figure 6)

Patients appear to have more in common with payers than physicians. Why?

- Payers are driven by costs, both direct and indirect (associated with managing adverse events associated with a therapy)
- Patients tend to bear the burden of both the direct (through co-pay levels) and indirect (through cost of managing adverse events)

Linked models address these disconnects directly by including them through feedback loops into the physician model

Payers appear to have more impact on physicians than patient requests so…

- If your client is short on research $$$, make sure you go for payers first then patients

Implication for client's basing launch strategy on physician-only research?: ***Caveat emptor!***

MODIFYING BAYESIAN NETWORKS FOR KEY DRIVERS ANALYSIS: AN OVERVIEW OF PRACTICAL IMPROVEMENTS

MIKE EGNER,
ROBERT A. HART, JR.
IPSOS SCIENCE CENTER

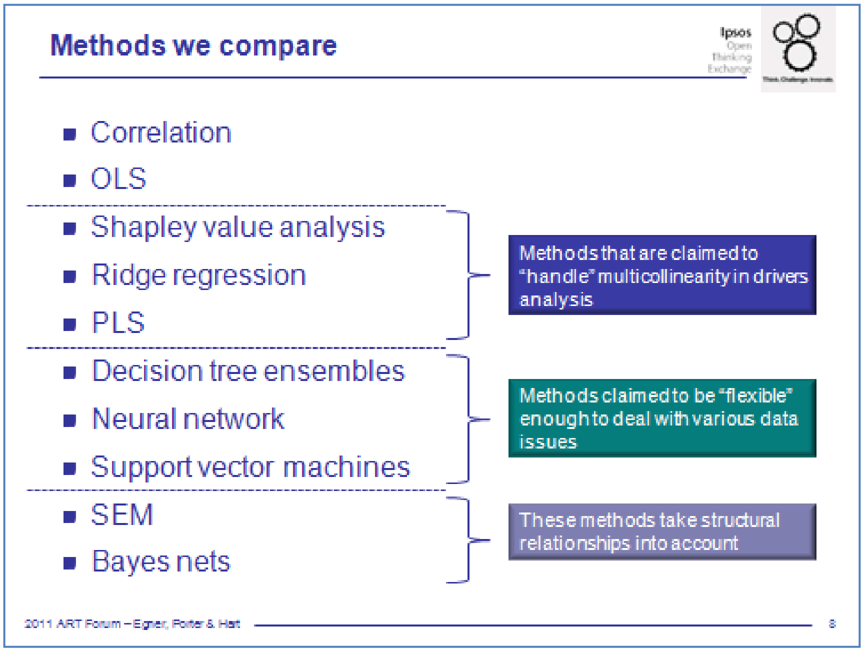
EXECUTIVE SUMMARY

Last year we investigated the relative effectiveness of methods commonly used to identify “key drivers” in marketing research (often using customer satisfaction, brand equity or other attitudinal survey measures). We discovered that Bayesian Networks outperformed all of the other methods tested when the data are multicollinear and/or contain non-linear relationships. Several modifications, however, were required to achieve these results and in this paper we detail the changes we made. We also discuss the problems that motivated the improvements: (1) small cell frequencies in the conditional probability tables, (2) a focus on beliefs rather than causal impacts, (3) uncertainty in a structural search, and (4) pseudo-simulation by selecting subsets of respondents. Understanding these hazard areas is critical for users and consumers of Bayesian Networks, and with the improvements we see great promise for Bayes Nets moving forward.

INTRODUCTION

In a previous paper we systematically compared the performance of methods used to estimate key drivers in market research (Egner, Porter and Hart 2011). We compared methods using a simple, practical metric: the methods' ability to accurately capture the true impact of a change in a key driver on the outcome variable. The authors compared several methods: some broadly used, some less common and a handful of techniques essentially designed to deal with problems commonly seen in survey data. The methods are listed below in Figure 1.

Figure 1: Methods Compared in Previous Drivers Research



Methods we compare

Ipsos Open Thinking Exchange

- Correlation
- OLS
- Shapley value analysis
- Ridge regression
- PLS
- Decision tree ensembles
- Neural network
- Support vector machines
- SEM
- Bayes nets

Methods that are claimed to "handle" multicollinearity in drivers analysis

Methods claimed to be "flexible" enough to deal with various data issues

These methods take structural relationships into account

2011 ART Forum – Egner, Porter & Hart 8

This research focused on two specific data characteristics that we worry about with market research survey data: multicollinearity and non-linearity. Non-linearity in the relationships between variables is (anecdotally at least) a less common issue with the data we use, and is also an easier problem to deal with using simple adaptations of methods like regression, or methods that inherently embrace such relationships. Multicollinearity, on the other hand, is a complex and near-ubiquitous problem in market research and needs a broader discussion.

Multicollinearity

The presence of multicollinearity is such a common problem with market research survey data that practitioners can now choose from a wide variety of methods claiming to treat the problem as part of the modeling process. The glaring nature of the problem – particularly when two similar drivers are given hugely different or even reverse-signed coefficients – has motivated many creative marketing scientists to try their hand at both developing new methods and comparing those methods that already exist.

The majority of the research on measures of relative importance use 'real-world' data sets from many different business cases (Soofi, Retzer and Yasai-Ardekani 2000; Pratt 1987; Soofi and Retzer 1995; Kruskal 1987). This clearly has one large (and ultimately fatal) flaw: none of us know what the *true* model looks like, and therefore there is no basis for evaluating methods in terms of their ability to capture and reflect reality.

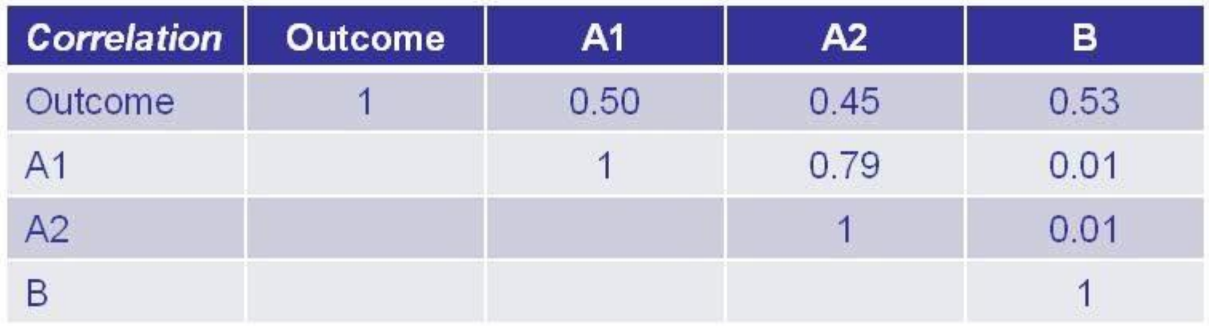
Some have defined a 'true' state of the world and generated simulated data from a pre-defined and known data generating process (see, e.g., Gromping 2009). This is clearly the best way to test a model's ability to capture the parameters defining the data. But there is a peculiar tendency in studies of this sort: they define the correlation structure amongst the predictor variables using the correlation matrix alone. At first blush this seems reasonable and those of us with classic social science or statistics training were taught to think about multicollinearity in these terms. This turns out, however, to be another fatal flaw that prevents us from properly understanding what is happening when we have correlated variables.

As we began designing our study, we asked a simple question: where does multicollinearity come from? We were not satisfied that the correlation matrix was an early enough place to start, and as we thought about the question and started to produce answers we knew we were approaching this in a new way. In the literature, no one is really talking about where multicollinearity *comes from*; i.e., what is *causing* it. Even the Wikipedia entry for multicollinearity describes it as a "statistical phenomenon" and the sections of the entry are "Definition of", "Detection of", "Consequences of", and "Remedies of".[40] Thinking about where multicollinearity comes from and considering all of the processes that result in data characterized by multicollinearity leads to a powerful conclusion:

Multicollinearity is inherently a *causal* phenomenon.

Contrary to how the term is commonly defined, we argue that multicollinearity is *not* a statistical phenomenon but rather a causal phenomenon that has statistical *evidence* and statistical *consequences*. This distinction is not trivial. Consider the following example from Egner, Porter and Hart (2011). When we treat multicollinearity as a statistical phenomenon the correlation matrix is the natural manifestation, as is shown in Table 1.

Table 1: Hypothetical Set of Correlations (Drivers A1 and A2 Correlated)

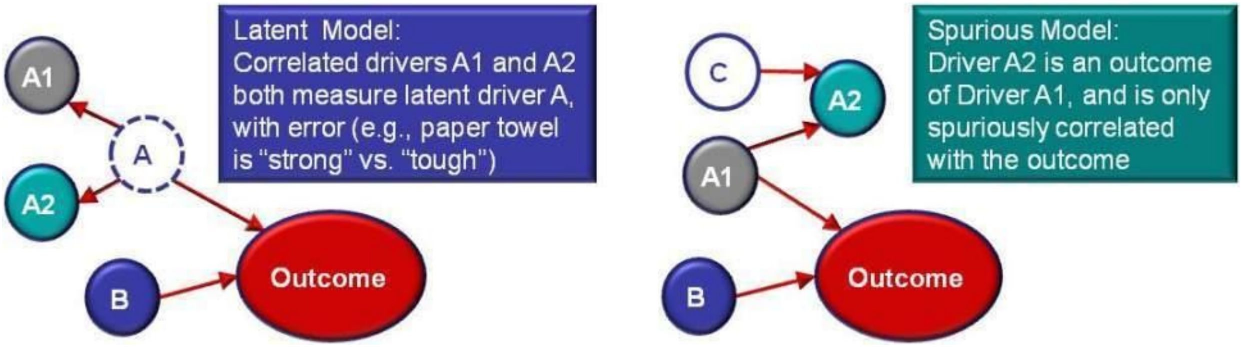


Correlation	Outcome	A1	A2	B
Outcome	1	0.50	0.45	0.53
A1		1	0.79	0.01
A2			1	0.01
B				1

Data can be simulated with these correlations and analyzed, but it's easy to show why this isn't enough information. Figure 2 shows two very different situations that can both give rise to the correlations shown above.

[40] http://en.wikipedia.org/wiki/Multicollinearity

Figure 2: Two Models with the Same Correlations


A1
A
A2
B
Outcome
Latent Model:
Correlated drivers A1 and A2 both measure latent driver A, with error (e.g., paper towel is "strong" vs. "tough")
C
A2
A1
B
Outcome
Spurious Model:
Driver A2 is an outcome of Driver A1, and is only spuriously correlated with the outcome

The diagram on the left shows a case where variables A1 and A2 are imperfect measures of an unobserved variable A (which impacts the Outcome directly), while variable B impacts the Outcome directly. On the right is an example where A1 and B both impact Outcome directly, and A2 is also impacted by A1 (and also by variable C). A2 has no relation to the Outcome other than as a spurious correlate.

There are a finite number of mechanisms that can produce a correlation matrix with non-zero elements off the main diagonal, such as (1) unobserved latent relationships, (2) spurious relationships, and (3) actual causal relationships between variables. Without knowing where multicollinearity comes from, it's impossible to tell if a method is properly capturing the relationships in the data, which has important managerial implications.

Egner, Porter and Hart (2011) show that in the presence of multicollinearity and/or non-linear effects, the only method that consistently avoided producing deeply misleading results was a modified Bayesian Network. This was largely due to the ability of Bayesian Networks to model the structural interrelationships *between* drivers, rather than either (1) assuming these relationships do not exist, or (2) arbitrarily splitting importance across correlated drivers, as other methods typically do.

However, Bayesian Networks as typically run are vulnerable to a number of shortcomings that can again deeply mislead a user interested in measuring the importance of key drivers. For the rest of the paper, we will describe the changes we made, highlighting the potential problems of standard Bayesian Networks in the context of key drivers, and offering concrete and practical fixes to these issues.

LIMITATIONS OF BAYESIAN BELIEF NETWORKS FOR DRIVERS

Small Cell Sizes in Conditional Probability Table

The first problem we consider when using Bayesian Networks is small cell sizes. A Bayesian Network is parameterized via conditional probabilities. For each unique combination of driver values (e.g., Driver A=3, Driver B=4, and Driver C=2), the actual distribution of the outcome variable among those respondents reporting that combination of driver inputs is measured and reported (the A,B,C triple 3,4,2 would be represented by a single row in the conditional probability table). Conditional probability is a powerful way to capture nonlinearities, synergies and other interactions between predictor variables and the outcome. However, in most practical applications many cells contain just a handful of observations (some have none at all). The reported distribution of the outcome variable is neither useful nor reliable when it is based on so few data points. Unfortunately this is not always apparent to the user, who must know what to look for. If ignored, there is a substantial risk that inferences can be based on the responses of very few individuals.

Belief Propagation

A second issue we address is the critical distinction between belief propagation and impact propagation. A Bayesian Network is typically referred to as a Bayesian *Belief* Network, which is entirely consistent with the typical usage of Bayes' Theorem – updating beliefs about something given the existence of some evidence. However, belief updating is indifferent to causal direction. Observing something yesterday will update my belief as to what will happen today; but by the same token, observing something today will update my belief as to what happened yesterday. The result of this causal indifference is that belief updating in a Bayesian Network – the typical way by which a user might evaluate what-if scenarios – allows both cause→effect and effect→cause changes that will skew measures of the practical importance of changing drivers. Furthermore, the process by which typical Bayesian Network approaches search for the structure among variables is itself indifferent to causal direction, and will make no attempts to distinguish between many A→B and B→A relationships. Again, this can lead to a deeply misleading understanding of "what effects what" for the purposes of understanding key drivers.

Structural Uncertainty

Structural search algorithms are not perfect, and at times there simply may not be enough information in the data to find a relationship between two variables or to assign causal direction to those that are found. In fact, the presence or absence of edges can change by including or omitting just a handful of observations. Small changes to the structure can have a large effect on estimates of the impact each predictor variable has on the outcome, yet nearly all programs find and present a single structure as "the" structure for modeling purposes. The uncertainty surrounding the specifics of the structure may be mentioned in passing, but are not incorporated into the estimates themselves.

Pseudo-Simulation

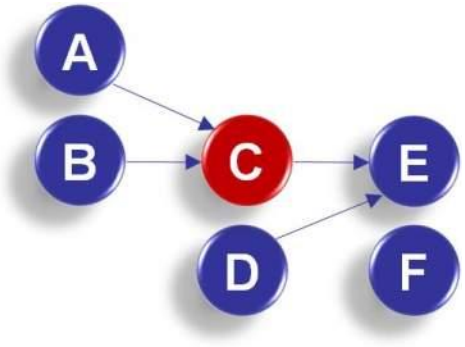
When simulating the impact of changes to a variable, the Bayes Net packages we've explored will typically just remove some respondents who scored poorly on that driver and report summary statistics for the remaining (higher-scoring) subset. Yet the poor scorers are precisely the people who we should be simulating a change on if we want to estimate the effect of improving this driver – and they may be fundamentally different from the respondents who already score high on the driver.

MODIFICATIONS/ADAPTATIONS

Modification 1: Tree-Based Cell Aggregation

To help illustrate the problems described above and provide intuition for the solutions we have developed, we simulated a simple dataset. Six variables (A,B,C,D,E,F), each on a 0-10 scale, have been generated based on the structure shown in Figure 3.

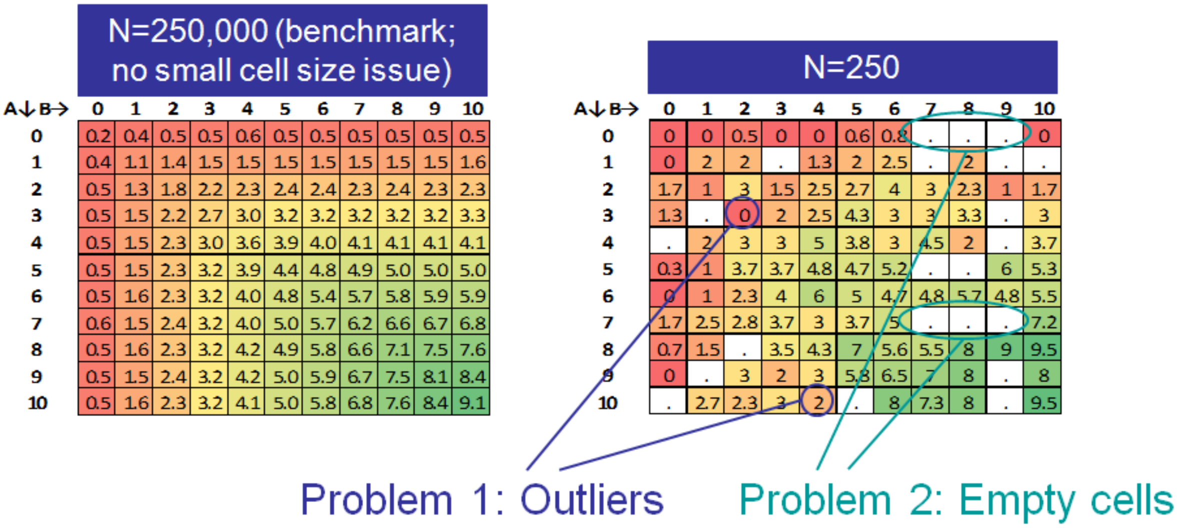
Figure 3: Structure of Simulated Data



In this hypothetical case, the user has been asked to run drivers using variable C as the outcome variable. Only two variables are actual causal inputs of C (variables A and B), and C is generated according to the relationship C = min(A,B) + ε. The simulated dataset has a sample size of N=250.

Typically, a Bayesian model will lay out a conditional probability table, but for a more birds-eye view we can aggregate this into a conditional *mean* table, showing the expected value of C at every possible combination of A and B. Figure 4 uses these tables to show how the small cell size problem can manifest itself. The table on the left shows the average values of C for each cell combination when n=250,000, which serves as a good benchmark. The table on the right shows an example of these same values when n=250.

Figure 4: Conditional Means of C when N=250,000 and N=250



N=250,000 (benchmark; no small cell size issue)

A↓ B→	0	1	2	3	4	5	6	7	8	9	10
0	0.2	0.4	0.5	0.5	0.6	0.5	0.5	0.5	0.5	0.5	0.5
1	0.4	1.1	1.4	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.6
2	0.5	1.3	1.8	2.2	2.3	2.4	2.4	2.3	2.4	2.3	2.3
3	0.5	1.5	2.2	2.7	3.0	3.2	3.2	3.2	3.2	3.2	3.3
4	0.5	1.5	2.3	3.0	3.6	3.9	4.0	4.1	4.1	4.1	4.1
5	0.5	1.5	2.3	3.2	3.9	4.4	4.8	4.9	5.0	5.0	5.0
6	0.5	1.6	2.3	3.2	4.0	4.8	5.4	5.7	5.8	5.9	5.9
7	0.6	1.5	2.4	3.2	4.0	5.0	5.7	6.2	6.6	6.7	6.8
8	0.5	1.6	2.3	3.2	4.2	4.9	5.8	6.6	7.1	7.5	7.6
9	0.5	1.5	2.4	3.2	4.2	5.0	5.9	6.7	7.5	8.1	8.4
10	0.5	1.6	2.3	3.2	4.1	5.0	5.8	6.8	7.6	8.4	9.1

N=250

A↓ B→	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0.5	0	0	0.6	0.8	.	.	.	0
1	0	2	2	.	1.3	2	2.5	.	2	.	.
2	1.7	1	3	1.5	2.5	2.7	4	3	2.3	1	1.7
3	1.3	.	0	2	2.5	4.3	3	3	3.3	.	3
4	.	2	3	3	5	3.8	3	4.5	2	.	3.7
5	0.3	1	3.7	3.7	4.8	4.7	5.2	.	.	6	5.3
6	0	1	2.3	4	6	5	4.7	4.8	5.7	4.8	5.5
7	1.7	2.5	2.8	3.7	3	3.7	5	.	.	.	7.2
8	0.7	1.5	.	3.5	4.3	7	5.6	5.5	8	9	9.5
9	0	.	3	2	3	5.8	6.5	7	8	.	8
10	.	2.7	2.3	3	2	.	8	7.3	8	.	9.5

Problem 1: Outliers

Problem 2: Empty cells

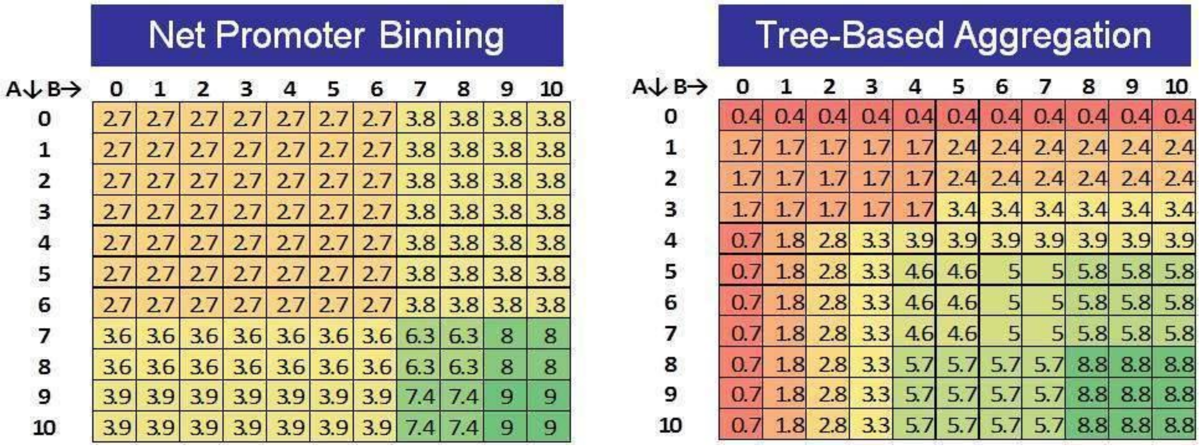
When the overall sample size is small, there are both cell entries that do not reflect the true underlying relationship very well and also many cells that have no entries at all.

To address the problem of small cell sizes in the conditional probability table practitioners can discretize, bin or even dichotomize variables in order to reduce the number of possible input combinations. While this reduces the cell-size problem it also runs the risk of washing out important thresholds in the data (and may still not eliminate all small cells).

To avoid these consequences we replaced the standard full disaggregation approach with a hybrid classification and regression tree (CART) approach, which disaggregates variables into the combinations that are most homogenous with respect to an outcome variable up to the point that a minimum cell size is reached. This eliminates the wild swings in conditional probabilities we see in a table afflicted by small cell sizes, and in fact makes the conditional probability table itself a far more useful tool for understanding the combinations that truly matter.

Figure 5 compares the effects of binning using two different binning rules. The table on the left shows a “Net-Promoter” style binning where the 11 point scale is trichotomized into low, medium and high groups (0-6, 7-8, 9-10). The table on the right shows the result of a tree-based approach.

Figure 5: “Net-Promoter” Type v Tree-Based Aggregation



Net Promoter Binning

A↓ B→	0	1	2	3	4	5	6	7	8	9	10
0	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
1	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
2	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
3	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
4	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
5	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
6	2.7	2.7	2.7	2.7	2.7	2.7	2.7	3.8	3.8	3.8	3.8
7	3.6	3.6	3.6	3.6	3.6	3.6	3.6	6.3	6.3	8	8
8	3.6	3.6	3.6	3.6	3.6	3.6	3.6	6.3	6.3	8	8
9	3.9	3.9	3.9	3.9	3.9	3.9	3.9	7.4	7.4	9	9
10	3.9	3.9	3.9	3.9	3.9	3.9	3.9	7.4	7.4	9	9

Tree-Based Aggregation

A↓ B→	0	1	2	3	4	5	6	7	8	9	10
0	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
1	1.7	1.7	1.7	1.7	1.7	2.4	2.4	2.4	2.4	2.4	2.4
2	1.7	1.7	1.7	1.7	1.7	2.4	2.4	2.4	2.4	2.4	2.4
3	1.7	1.7	1.7	1.7	1.7	3.4	3.4	3.4	3.4	3.4	3.4
4	0.7	1.8	2.8	3.3	3.9	3.9	3.9	3.9	3.9	3.9	3.9
5	0.7	1.8	2.8	3.3	4.6	4.6	5	5	5.8	5.8	5.8
6	0.7	1.8	2.8	3.3	4.6	4.6	5	5	5.8	5.8	5.8
7	0.7	1.8	2.8	3.3	4.6	4.6	5	5	5.8	5.8	5.8
8	0.7	1.8	2.8	3.3	5.7	5.7	5.7	5.7	8.8	8.8	8.8
9	0.7	1.8	2.8	3.3	5.7	5.7	5.7	5.7	8.8	8.8	8.8
10	0.7	1.8	2.8	3.3	5.7	5.7	5.7	5.7	8.8	8.8	8.8

Compare with the benchmark on the earlier slide: tree-based aggregation works off of the data, automatically capturing important thresholds (and ignoring unimportant ones). Significantly, using a method like this essentially maximizes the information available but doesn’t promise more than is truly there.

Modification 2: Causal Search Algorithms

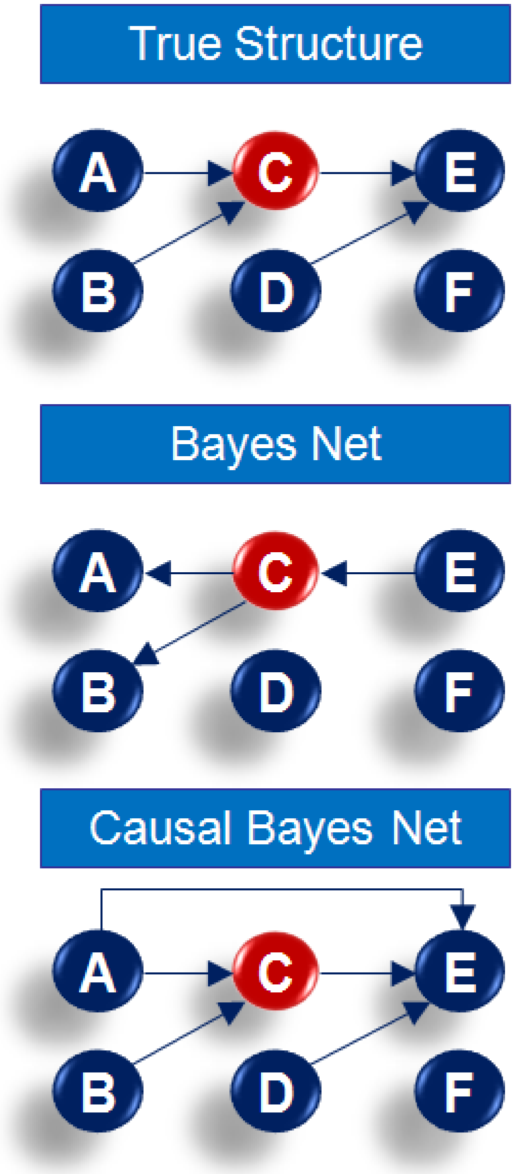
The purpose of identifying key drivers is to provide managerial guidance on actions that will help produce a positive change in some aspect of the business. Since this goal is clearly *causal* in nature we must evolve from the causally indifferent world of belief updates/propagation into the world of cause and effect. We replace belief propagation with *impact* propagation by adopting the causal structural search algorithms developed by Judea Pearl and the scholars at the Carnegie Mellon TETRAD Project (Pearl 2000, 2009, 2010; Spirtes et al. 1993). This approach is used to identify the direction of causal relationships between drivers using conditional independence logic, and further we limit impact propagation to be strictly “downhill” in that updates can only flow from causes to effects, and not vice versa.

The basic premise of the conditional independence approach is that induced dependence helps orient causal directions among related variables. For example, if variables A and B are zero-order independent from each other, but (1) are both dependent on variable C and (2) become conditionally dependent on each other once you control for C, then you can orient the relationships as A→C and B→C.

What does this mean, exactly? Using our example, we have explicitly simulated A and B to be independent: knowing the value of one tells you absolutely nothing about the value of the other. However, this changes once you control for C. Given that C=min(A,B) (ignoring measurement error for the moment), knowing the value of A at a particular level of C suddenly *does* tell you something about the value of B. For example, at C=3, knowing that A=6 tells you that B must equal 3 (since C is simply the minimum of A and B). This pattern of dependencies implies that A and B are causal inputs of C. With any other orientation of the causal arrows, we would have expected to see a zero-order relationship between variables A and B, not merely one that only popped up conditional on variable C.

Feeding the previously-discussed dataset into a traditional Bayes Net package and one modified to search for causal structure produced the results shown in Figure 6.

Figure 6: True and Estimated Causal Structures



Comparing the two lower structures with the top (true) structure, both programs mostly find the right connections (traditional BN misses the relationship between D and E, while the causal BN adds an unnecessary link between A and E), but the traditional Bayes Net has reversed the key orientations. This is reasonable from a belief-updating

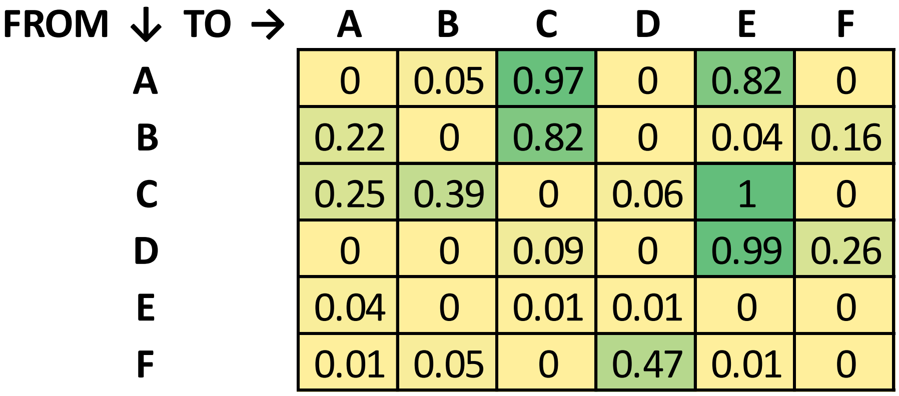
POV – observing an effect updates your belief about whether the cause occurred – but it is not useful for drivers analysis.

Modification 3: Bootstrapped Structures

As noted earlier, given the inherent uncertainty in any modeling procedure, we see risk in settling on a single structure for the purposes of a drivers analysis. To address this, we bootstrap the data multiple times (typically 100), estimate a distinct structure from each of these bootstraps, run our analysis on each structure, and average the results across the structures.

As a result, if a particular structural component could really “go either way,” we will have some bootstraps with it and some without, and can then average driver importance scores over the uncertainty. Figure 7 shows the percentage of times that each edge appears in the simulated dataset we have been using in the previous examples.

Figure 7: Fraction of Bootstraps Where Edge Appears



FROM ↓ TO →	A	B	C	D	E	F
A	0	0.05	0.97	0	0.82	0
B	0.22	0	0.82	0	0.04	0.16
C	0.25	0.39	0	0.06	1	0
D	0	0	0.09	0	0.99	0.26
E	0.04	0	0.01	0.01	0	0
F	0.01	0.05	0	0.47	0.01	0

This approach both respects the uncertainty that exists in the data, and actually provides important information on the degree to which we can be confident that a driver actually is a driver. Note that we will sometimes find bi-directed edges, so the sums across a dyad (e.g. B→C and C→B) can exceed one.

This is perhaps our simplest modification, but it may be the most important. Given the processing power of today’s computers, there is little reason to stake driver importance scores (and the business decisions that rely on them) on an all-or-nothing bet that a particular link in a structure definitely is (or is not) present.

Modification 4: Actual Simulation

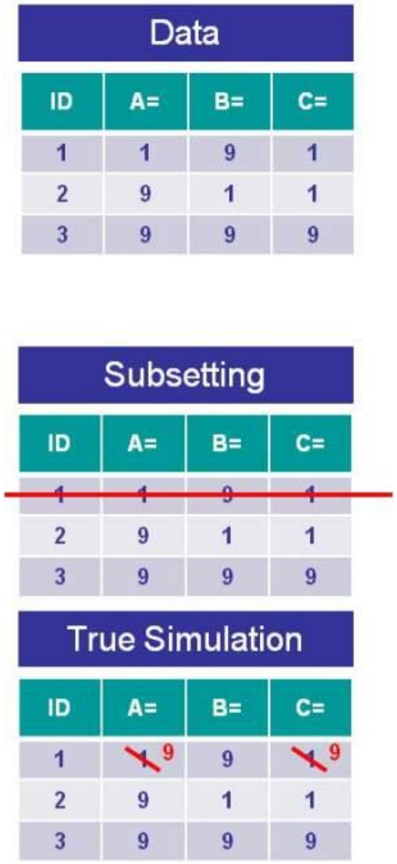
Perhaps the key lesson of Bayes’ theorem is that it is important to get the right denominator when calculating probabilities. Thinking back to the classic example of the theorem – the highly accurate test for a disease that is only present in a small fraction of the population – the key insight of Bayesian thinking is that one must first subset down to the population of all individuals with a positive test result, and only then look for the prevalence of the disease among that subgroup. If the disease is rare enough, the false positives may swamp the true positives among this subgroup, leading to the counterintuitive insight that this highly accurate test, when positive, ends up usually being wrong.

This is obviously a powerful approach. However, in the context of simulation and "what if" scenarios, a straight application of this approach can actually lead to trivial results. Specifically, asking the question "What would happen if everyone moved to top box on Driver X?" becomes simply "What is the current value of the outcome among the subset of respondents already at top box on Driver X?"

This logic is usually carried over into the simulation tools in Bayes Net packages: if you try to "simulate" an increase in the driver, the program will typically just remove some respondents who scored poorly on that driver and report summary statistics for the remaining (higher-scoring) subset. Yet the poor scorers are precisely the people who we should be simulating a change on, if we want to estimate the effect of improving this driver – and they may be fundamentally different from the respondents who are already scoring well on the driver.

Figure 8, below, illustrates the differences between simulation and subsetting on three rows of data when answering the following "what if" question: what would happen if everyone moved to variable A=9?

Figure 8: Subsetting Versus Simulation



Data

ID	A=	B=	C=
1	1	9	1
2	9	1	1
3	9	9	9

Subsetting

ID	A=	B=	C=
~~1~~	~~1~~	~~9~~	~~1~~
2	9	1	1
3	9	9	9

True Simulation

ID	A=	B=	C=
1	~~1~~ 9	9	~~1~~ 9
2	9	1	1
3	9	9	9

Under the subsetting approach (middle table), we would simply remove the rows where variable A is currently not equal to nine. Under the simulation approach (lower table) we would (1) preserve all rows, (2) change the non-nine values of variable A into nines, and (3) estimate a new value for the outcome variable, based on the relationships we have already modeled. In this case, since we have captured that C is the minimum of A and B, we would peg the adjusted outcome value at nine as well (or something close to it, depending on measurement error).

Put another way, subsetting yields a vector of original outcome values with length equal to some (potentially very small) subset of the original sample size, while simulation yields a vector of *adjusted* outcome values of length equal to the *original* sample size. To the extent that those respondents who are not currently at the "what if" value of the driver differ from those who are, this will yield different results (as in Figure 8, where the average outcome in this "what if" scenario is 5 under subsetting and 6.3 under simulation).

Rather than subset, our calculations take the simulation approach. We estimate new output values for all respondents – based on the new area of the conditional probability table they are "jumping into" – and take an average of these adjusted values across all rows of the data.

CONCLUSION

Once certain steps are taken to address the potential problem areas with existing Bayesian network software, then this approach to modeling drivers of marketing attitudes and behavior has a tremendous amount of promise. By acknowledging structure and seeking to understand it, we gain a richer understanding of both the business situation we're exploring and the data we've collected to do so. There is a tremendous amount of future research to be done along this path, including understanding of how this approach can be further improved and aligned with other work tackling similar or related issues (see, e.g. Büschken, Otter and Allenby 2011). Embracing the reality of causal structure more explicitly can also help us all become better researchers as we are forced to do what we should be doing already: thinking critically about the causal relationships that exist in the marketplace.

BIBLIOGRAPHY

Achen, C.H. (1982). *Interpreting and using regression.* Beverly Hills, CA: Sage.

Büschken, J., Otter, T. and Allenby, G. (2011). Do we halo or form? A Bayesian mixture model for customer satisfaction data. Presented at American Marketing Association ART Forum, Palm Desert, CA.

Egner, M., Porter, S. and Hart, R. (2011). Key drivers methods in market research: a comparative analysis. Presented at American Marketing Association ART Forum, Palm Desert, CA.

Gromping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician* 63: 308-319.

Johnson, J. W., and LeBreton, J. M. (2004). History and use of relative importance indices in organizational research. *Organizational Research Methods* 7: 238-257.

Kruskal, W. (1987). Relative importance by averaging over orderings. *The American Statistician* 41: 6-10.

Kruskal, W., and Majors, R. (1989). Concepts of relative importance in scientific literature. *The American Statistician* 43: 2-6.

LeBreton, J. M., Ployhart, R. E., and Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods* 7: 258-282.

Pearl, J. (2000, 2009). *Causality: Models, Reasoning and Inference.* New York: Cambridge University Press.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics* 6(2): Article 7.

Pratt, J. W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained. In Pukkila, T. and Puntanen, S., eds., *Proceedings of the Second International Tampere Conference in Statistics*, University of Tampere, Finland, pp. 245-260.

Soofi, E. S., Retzer, J. J., and Yasai-Ardekani, M. (2000). A framework for measuring the importance of variables with applications to management research and decision models. *Decision Sciences Journal* 31(3): 596-625.

Soofi, E. S. and Retzer, J. J. (1995). A review of relative importance measures in statistics. In *The ASA Proceedings of Bayesian Statistical Science Section,* American Statistical Association, pp. 66-70.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, prediction and search*. New York: Springer-Verlag.

VOLUMETRIC CONJOINT ANALYSIS INCORPORATING FIXED COSTS

JOHN R. HOWELL
GREG M. ALLENBY
OHIO STATE UNIVERSITY

Conjoint analysis has been used in many different situations to investigate feature and price configurations that appeal to customers. Many of the advances in recent years have focused on allowing the product features to be flexibly designed to accurately reflect the products available in the marketplace. However since the popularization of CBC in the early 90's, little has been done in practice to address the common marketing question of how much of products consumers purchase. There are many common situations where consumers purchase multiple units of a product instead of simply choosing one from a set as is assumed in multinomial choice models. Some recent work in has highlighted the need for these models. Thomas Eagle (2010) presented a paper at the previous Sawtooth Software Conference that highlights the current state of practice. He presented three common methods for modeling volumetric choice and briefly highlighted a fourth method. The model presented in this paper is an extension of the economic models of choice mentioned, but not discussed in that paper.

When considering volumetric choices a number of additional issues arise that are not present in the discrete choice setting. In addition to considering the joint/continuous aspects of demand highlighted in Eagle (2010), complex pricing structures arise when considering volume purchases. Pricing strategies such as buy-one-get-one-free or bundle discounts are commonly observed, but only possible to consider in a volumetric setting as the per unit pricing is not constant. This paper demonstrates how to use an economic modeling framework we have called Volumetric Conjoint Analysis to investigate the inclusion of fixed costs into a pricing strategy. It also compares it to the most common existing method for handling this type of data.

Fixed costs can arise in many different settings and can even vary from person to person for the same product category. Examples of products with fixed costs include things like printers and ink, video game consoles and video games, country club or gym memberships, discount clubs like Costco or Sam's Club, and vacations. If you take a slightly broader view of fixed costs and consider costs that are not monetary or explicitly incurred almost all situations that have a repeat or volume purchase decision have an aspect of fixed cost. This includes things like the time and mental effort that is required to become familiar with a specific category, the time and distance travelled to shop at a grocery store as in Bell et al. (1998), or the cost associated with learning a hobby such as golf or skiing. One thing about fixed costs is that they often exist at multiple levels in the hierarchy of decision-making and are the most relevant cost when making category participation decisions.

Fixed costs can be considered an extreme form of volume discounting since the fixed cost is generally incurred with the first unit purchased, but is not present for additional unit purchases. Rational consumers must therefore anticipate their volume purchases when making a decision whether to participate in a product category as the total cost of participation needs to fit within their budget constraint. When fixed costs are high this naturally leads to a gap in observed purchase volumes just above zero. This is due to the average cost of purchasing a product being exceptionally high for small volumes, but quickly decreasing as the fixed cost is distributed across higher volumes of purchases.

Fixed costs are associated with a number of behavioral phenomena. Fixed costs can lead to local monopolies. Local monopolies arise when one firm has pricing leverage over a limited set of customers. This means that companies can increase prices above the market competitive rate and not experience significant switching. Local monopolies are present when switching costs are high. The switching costs impose an implicit fixed cost on competing brands. This means that consumers are required to pay both the explicit cost for the desired good as well as the implicit cost associated with switching. The additional cost reduces the quantity they are able to purchase compared to the existing brand and thus their overall utility.

Fixed costs also lead to a natural discrimination between high volume and low volume customers. In the face of fixed costs low volume consumers will naturally self-select out of the market because the average cost is unreasonably high. In any marketing situation some customers are more profitable than others. Firms with limited resources desire to focus on the most profitable consumers. Because fixed costs lead to screening out low volume purchasers, customers self-identify as high volume, more profitable purchasers.

Since the low volume customers natural self-select out of the market the data will show many zeros, but very few small volume purchases. This is a natural outgrowth of market structure and not externally imposed. Figure 1 below demonstrates this phenomenon. The figure is plotted from a simulated three-good market. The plot projects the three-dimensional purchases down onto two dimensions and has been colored to highlight the different groups. The data were generated by randomly drawing a preference parameter for each observation. The observations were assumed to have a homogenous budget and face identical marketplace prices. The preference parameters along with the budget and fixed and variable prices were then used to simulate purchase decisions according to an economic utility maximization model. The most obvious feature of the data is the lack of points between zero purchases and positive quantities. This gap is due to the nature of fixed costs. The higher marginal utility for the unpurchased good does not compensate for the decrease in purchasing power due the fixed costs. Customers would therefore prefer to spend their money on existing products rather than invest in a new product unless the expected utility of the new product exceeds the fixed cost of that product.

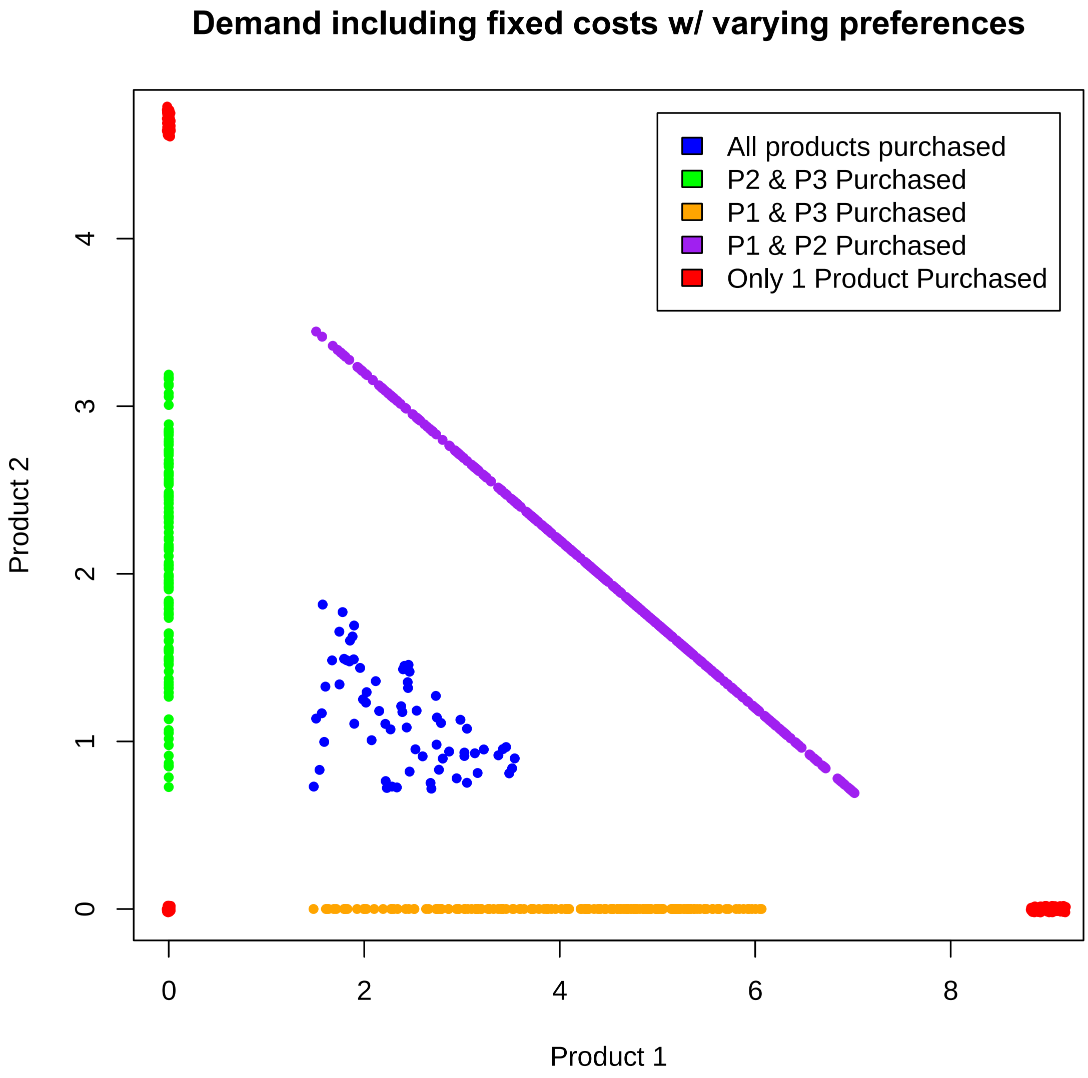


Figure 7. Simulated demand with the presence of fixed costs. Data simulated using homogenous prices and budgets, but random preferences

Volumetric Models

As previously stated examining fixed costs necessitates a volumetric model. In a discrete choice model it is impossible to separate the fixed and variable costs as both costs are only incurred once. Existing volumetric models do not explicitly handle the unique nature of fixed costs. The model proposed in this paper extends standard economic models to include a fixed cost component.

We recognize that these models may be unfamiliar to many readers. Our personal experience consulting with marketing research practitioners suggests that the most common method for estimating volume follows a choice modeling approach where the choice model is fit assuming the respondent makes a series of independent choices. In simulation the model must then be reweighted to reflect the actual volume the respondent is expected to purchase. For a complete description of the method and the problems associated with it see Eagle (2010).

Some of these problems of using the choice modeling approach include the inability to make proper inference from these models, a lack of confidence in the predictions made, the inability to properly account for relevant quantities such as cross-price elasticities and the confounding of preference and volume. Because these models are unfamiliar we will provide a brief overview of the basic model before extending it to include the fixed cost component. For a more complete treatment of these models please see Chandulkala et al. (2007)

VOLUMETRIC CONJOINT ANALYSIS

Volumetric Conjoint Analysis is based on a foundational principle that consumers choose a set of purchases in order to maximize their utility subject to a budget constraint. In mathematics this is represented by a utility function and a budget constraint. This can be written as:

$$\begin{aligned} &\max U(x) \\ &\text{s.t.}\ \ E \geq p \cdot x \end{aligned}$$

where U(x) is a utility function, E is the budget, p is the prices, and x represents the vector of quantities purchased. The concept should be familiar to CBC users where the assumption is that survey respondents choose the concept that has the highest utility. The difference is that in CBC the price is assumed to be a product feature instead of a separate constraint. Economic models instead treat prices as unique attributes that do not themselves have an associated utility. They assume that consumers do not have any inherent utility for money, but simply use it as a means of easing the exchange of goods between economic agents. This specification cleanly separates the costs or what consumers give up from the benefits or what they receive. Formulating the utility function in this way ensures that price sensitivities have the correct sign and form without the use of external constraints.

Economic utility functions can take many forms and the preferred formulation is often problem specific, but we have employed the following form with considerable success:

$$U(X) = \sum_{k=1}^{K} \frac{\Psi_k}{\gamma} \log(\gamma x_k + 1) + \log(z)$$

where k indexes the goods in consideration, x_k is the quantity of good k purchased, z represents the amount of the budget allocated to the none option, Ψ_k is the baseline utility for good k, and γ is a scaling parameter that controls for both the rate that consumers satiate and appropriately scales the quantities. There are a number of important features in the utility function. The log structure ensures that the utility exhibits decreasing marginal returns and allows for multiple products to have positive demand. The + 1 in the log function ensures that products can have 0 purchases. When an outside good or none option is included in the utility specification the budget constraint is guaranteed to bind allowing us to replace the greater than or equal sign in the budget constraint with an equality sign. With this utility function the entire function becomes:

$$\max_x U(x) = \sum_{k=1}^{K} \frac{\Psi_k}{\gamma} \log(\gamma x_k + 1) + \log(z)$$

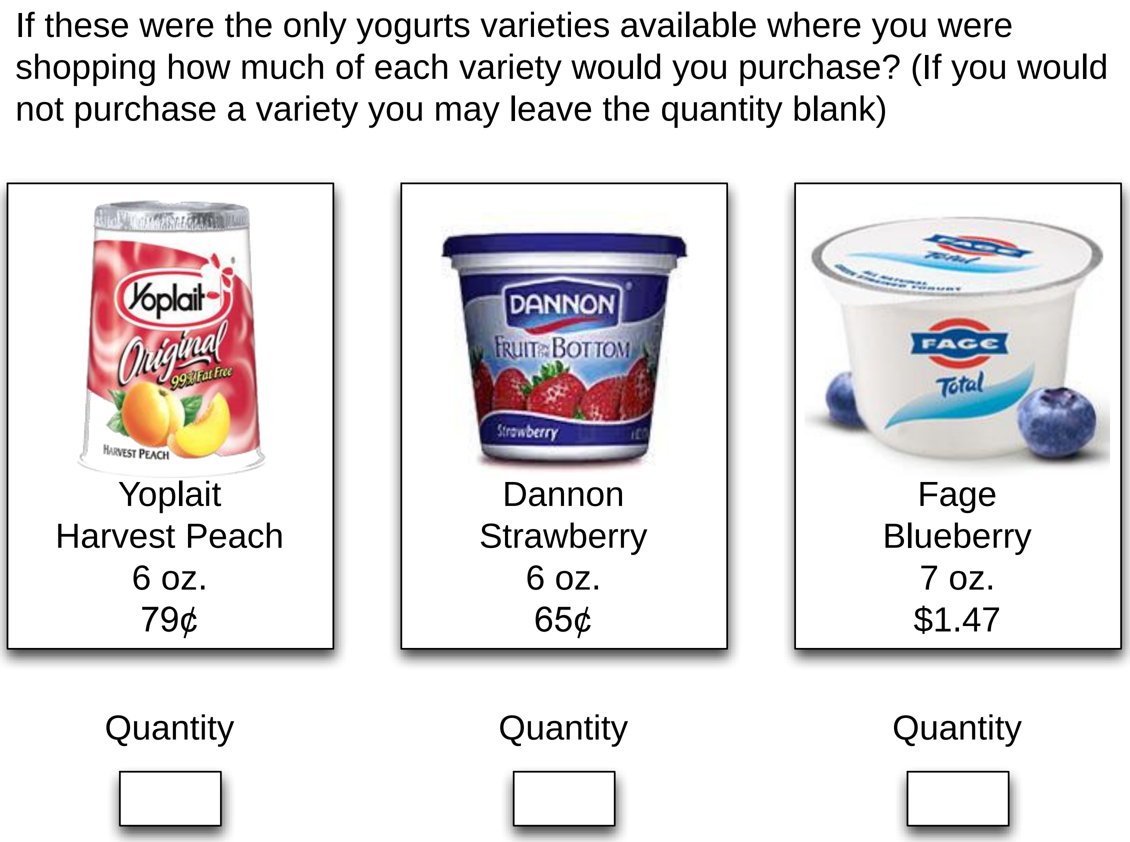
$$\text{s.t. } E = p \cdot x + z$$

The Ψ_k can be further parameterized by using the attributes and levels and part-worth utilities as in:

$$\Psi_k = e^{a_{k1}\beta_1 + a_{k2}\beta_2 + \ldots + a_{kl}\beta_l + \varepsilon_k}$$

where a_{kl} is the attribute level for the lth attribute for the kth concept.

A concrete example may make things easier to understand. Consider the conjoint task in Figure 2. Respondents are asked to specify how many yogurts of each flavor they would purchase. This is a typical volumetric conjoint task that can be easily programmed in Sawtooth Software's CBC/Web program.



If these were the only yogurts varieties available where you were shopping how much of each variety would you purchase? (If you would not purchase a variety you may leave the quantity blank)

Yoplait Harvest Peach 6 oz. 79¢	Dannon Strawberry 6 oz. 65¢	Fage Blueberry 7 oz. $1.47
Quantity □	Quantity □	Quantity □

Figure 2. Example Volumetric Conjoint Analysis Task

For this example there are three products in our choice set with three attributes plus price for each one. Note that there is not an explicit fixed cost in this example. We will introduce the fixed cost later.

The utility function for this example would be:

$$U(x_Y, x_D, x_F, z) = \frac{\Psi_Y}{\gamma} log(\gamma x_Y + 1) + \frac{\Psi_D}{\gamma} log(\gamma x_D + 1) + \frac{\Psi_F}{\gamma} log(\gamma x_F + 1) + log(z)$$
$$\text{s.t. } E = p_Y x_Y + p_D x_D + p_F x_F + z$$

where the subscripts refer to the specific brands. Specifically:

$$\Psi_Y = \beta_{YOPLAIT} + \beta_{PEACH} + \beta_{6OZ} + \epsilon_Y$$

Using a technique called the Karush-Kuhn-Tucker conditions allows you to solve this constrained optimization problem. The full statistical specification is beyond the scope of this paper, but can be found in the Chandukala et. al. (2007) paper.

Previous results have showed significant improvements over existing models and the estimation is straightforward (Satomura et al. 2011). If we assume extreme value errors, we can compute a closed form expression for the likelihood of the data and use Metropolis-Hastings to compute the posterior probabilities. Unfortunately these routines have not made their way into packaged software at this time making the models more difficult to fit. The part-worth utilities, the γ parameter, and the budget constraint E can all be estimated from the conjoint data. Once the parameters are calculated they can then be used to simulate a wide variety of outcomes similar to the way a traditional choice simulator functions.

VOLUMETRIC CONJOINT ANALYSIS WITH FIXED COSTS

Incorporating fixed costs into Volumetric Conjoint Analysis involves a simple extension to the previous model. The fixed cost represents an additional cost that enters into the utility maximization problem for those brands that are purchased. This can be represented as an additional indicator indicating that the target fixed cost had been incurred. The mathematical model for this is:

$$\max_x U(X) = \sum_{k=1}^{K} y_k \frac{\Psi_k}{\gamma} \log(\gamma x_k + 1) + \log(z)$$
$$\text{s.t. } E = \sum_{k=1}^{K} p_k x_k + \sum_{k=1}^{K} y_k c_k + z$$

where y_k is an indicator variable taking on a value of 1 if $x_k > 0$ and 0 otherwise and c_k is the fixed cost of the k^{th} product. The other variables retain their previous meaning. In this specific formulation the y_k variables in the utility function provide redundant information since even though the structure of the utility function and budget constraint are very similar to the previous

$$\log(\gamma x_k + 1) = 0 \text{ if } x_k = 0.$$

utility function, the indicator function for the fixed costs represent a unique challenge for solving the utility maximization problem. The Karush-Kuhn-Tucker conditions require the budget constraint to be continuously differentiable at least once and the indicator variable creates a discontinuity in the budget constraint.

To work around this problem we can leverage our knowledge of the utility function and budget constraint to divide the space spanned by the utility maximization problem into sub-regions where each sub-region is fully differentiable. We can then compare the utility maximizing solution for each of the sub-region in order to arrive at an overall utility maximizing solution. When translating this maximization strategy into a likelihood function however we are unable to solve for a tractable closed form expression, as the likelihood is highly irregular. We solve this problem by applying a Bayesian estimation technique called data augmentation. Similar methods have been employed in a number of previous papers where they are described in more detail. (Gilbride and Allenby 2004; Tanner and Wong 1987) The basic idea is that we assume respondents answer the conjoint questionnaire according to the utility maximization problem previously presented. This utility maximization could be easily solved if the Ψ_k parameters were known, but instead these are parameters that are stochastically determined. In order to solve the problem we first make a draw of the Ψ_k parameters. We then check to see if the draws are consistent with the utility maximizing solution. If the draw is consistent with the observed data, we use that draw in a Bayesian regression to calculate the part-worth estimates. Note that all this is done within a Gibbs sampler.

There are a couple of important points regarding the draws of Ψ_k. The first is the utility maximization step leads to a density contribution for interior solutions and a mass contribution for corner solutions. This leads to the draws for Ψ_k needing to be made from a truncated distribution for the cases where $x_k= 0$, but are known exactly conditional on E and γ. Fortunately these exact values can be calculated directly in closed form. The case where $x_k = 0$ is a little more challenging. The draws for these parameters are made from a truncated distribution, but the truncation points are dependent on E and γ and also cannot be computed in closed form. We employ a rejection sampling approach for these parameters by repeatedly drawing candidate draws from the untruncated distribution until we find a draw that satisfies the utility maximization solution. Because the rejection sampling routine is a relatively expensive operation we employ an adaptive-rejection sampling scheme. This allows us to use prior rejected draws to inform the algorithm of proper truncation points. This adaptive rejection sampling greatly speeds the rejection sampling process.

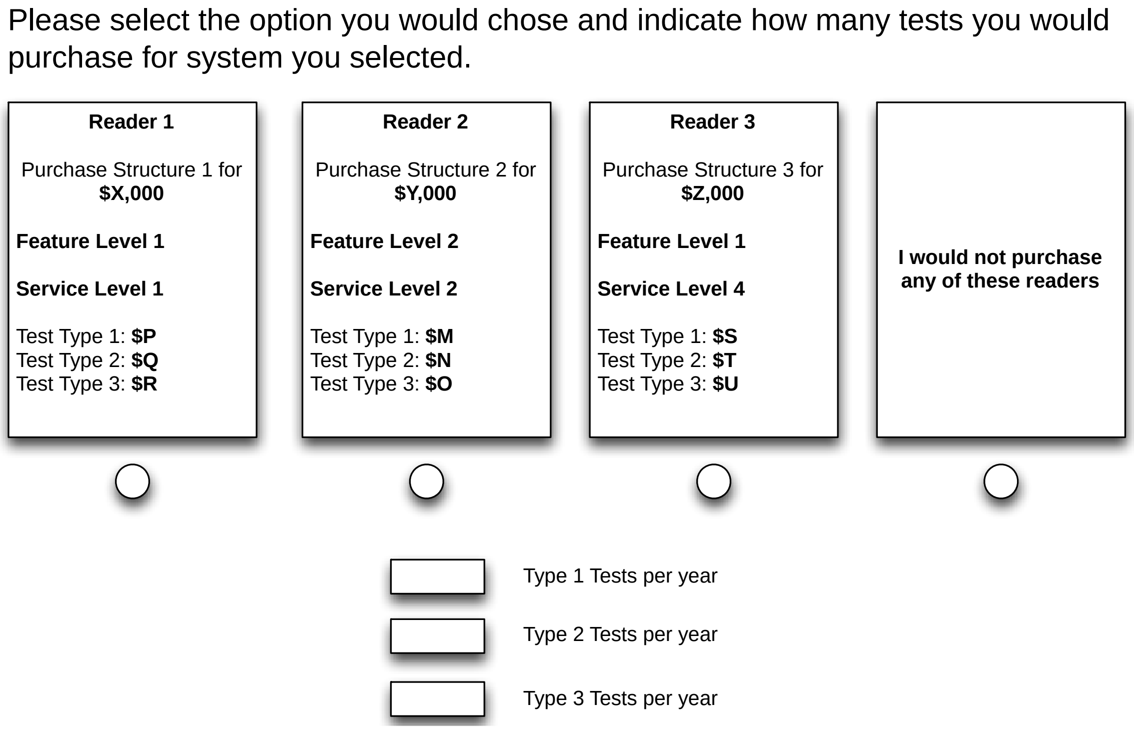
Simulated data shows that this data augmentation with adaptive rejection sampling approach converges quickly and can accurately recover the true parameters.

CASE STUDY

One of the authors' consulting clients approached them wishing to conduct a conjoint analysis involving fixed costs. The client graciously allowed us to present the findings, but wished to remain anonymous. We have therefore disguised the results to protect the anonymity of the client.

The product setting is an agricultural equipment manufacturer. This manufacturer has developed a new machine that is capable of detecting three different diseases. The farmer places a sample on a test strip and inserts the strip into the machine to run the test. Two separate components will be sold: the machine and the reader, and the client was interested in testing different pricing strategies for both. There are a total of three different test strips that will each be priced differently to test for the three different diseases. Additionally this product is new to the marketplace and does not currently have direct competitors. Therefore there is no historical data to benchmark the model against. It was determined that a volumetric conjoint analysis with fixed costs would be appropriate for testing this product.

An example screen for this product can be found in Figure 3. The actual screens were designed and programming using Sawtooth Software's CBC/Web program.



Please select the option you would chose and indicate how many tests you would purchase for system you selected.

Reader 1	Reader 2	Reader 3	
Purchase Structure 1 for **$X,000**	Purchase Structure 2 for **$Y,000**	Purchase Structure 3 for **$Z,000**	**I would not purchase any of these readers**
Feature Level 1	**Feature Level 2**	**Feature Level 1**	
Service Level 1	**Service Level 2**	**Service Level 4**	
Test Type 1: **$P** Test Type 2: **$Q** Test Type 3: **$R**	Test Type 1: **$M** Test Type 2: **$N** Test Type 3: **$O**	Test Type 1: **$S** Test Type 2: **$T** Test Type 3: **$U**	
○	○	○	○

□ Type 1 Tests per year

□ Type 2 Tests per year

□ Type 3 Tests per year

Figure 3. Example Volumetric Conjoint Analysis with Fixed Cost screen

The only addition to the standard discrete choice screen was the addition of three additional volume questions. It was determined that it would be too heavy of a burden to respondents if they were asked to provide volumes for each of the three readers so we employed a choose-then-elaborate strategy. Due to the nature of the readers and the price it is unlikely that a respondent would desire to purchase multiple readers. Here the reader purchase represents the fixed cost component of the purchase and the additional tests represent the variable cost component. Both the fixed and variable costs are required to use the machine. One minor difference between this study and the previous exercise is that each fixed cost component is associated with three separate cost components. The model handles this in a straightforward manner by reducing the number of y_k s from nine to three. It is then possible to collapse the model into choice between three separate sun-utility models. The estimation procedure then follows as previously described.

The data was collected over an 8-week period in fall of 2011 using a phone recruit to web strategy. Two hundred twenty respondents were collected out of a population of 3500 but only 147 were usable for estimation, as the remaining respondents chose no purchase for all the tasks.

This exercise proved to be difficult for respondents due to the volume questions and the newness of the product category. Respondents were required to project volumes for products they had not previously been exposed to. An unfortunate side effect is that approximately two-thirds of the 147 respondents alternated between zero and the single positive number for their volume for each task.

RMSE	Holdout 1	Holdout 2
Reader Volume	8.71	5.22
Reader Share	5.93%	3.55%

Table 1. Reader Prediction Errors (RMSE)

Earlier in this paper we showed that fixed costs would induce a gap between zero and the smallest positive quantity that respondents would choose. This allowed them to spread the high fixed cost over a large number of units, thus reducing the per unit cost. In this case the testing system only becomes economically reasonable if the number of tests is large. This is reflected in the actual data as the lowest non-zero number of tests is 60 reflecting the gap between 0 and positive participation in the market. The number of tests projected to be ordered ranged from 60 to 22500.

The conjoint exercise consisted of 14 tasks with 3 concepts each. Two of the conjoint tasks were held out for external validation purposes. Note that the comparisons to the holdout tasks may seem unfamiliar to regular CBC users. This is due to the fact that CBC is a share-based model so the natural comparison is difference in percentages shares between the model and the holdout tasks. The Volumetric Conjoint Analysis procedure is a volume-based task. To assist in the comparison we have converted the reader purchases into share predictions, but we are unable to convert the volume predictions into shares. Table 1 presents the share predictions for the readers. Figures 4 and 5 present a comparison of the predicting and actual volumes for the holdout tasks. This was calculated by summing the volumes for all three readers. Volume predictions are especially sensitive to the number of respondents predicted to purchase the products as the volumes for this study are regularly in the thousands for some respondents. A single respondent that is misclassified as purchasing or not purchasing can dramatically change the results.

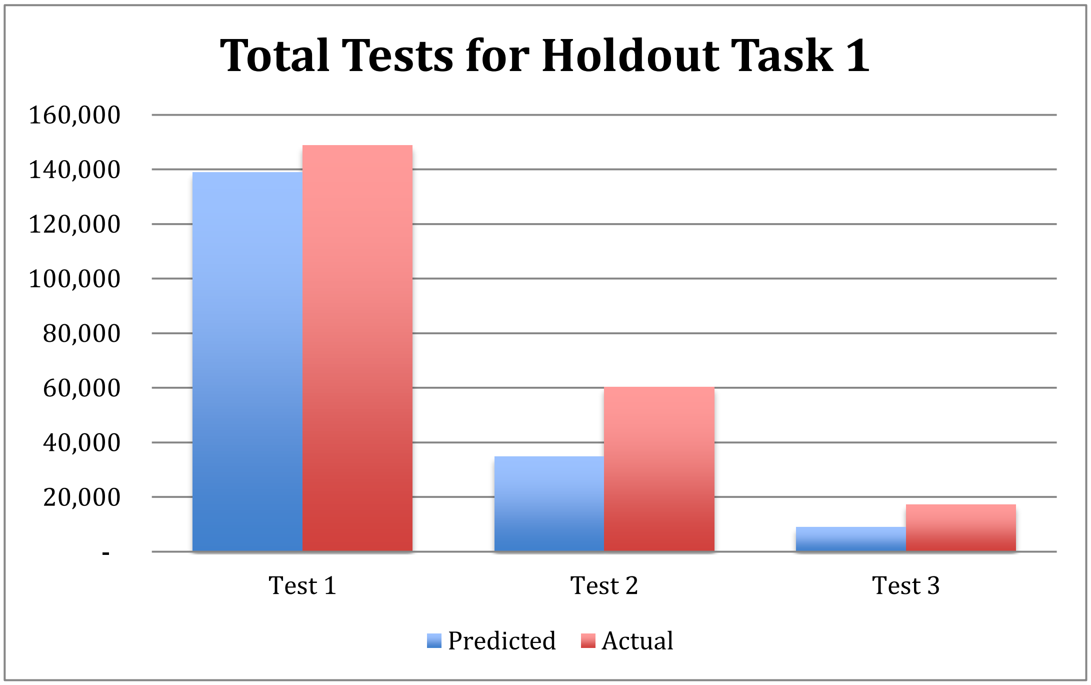


Figure 4. Test Volume Predictions for Holdout 1

We also ran a comparison to a CBC style volumetric study. There are a number of ways that have been proposed to handle volumetric studies using CBC. The method described in Orme (2010) is not practical for this study as respondents provided volume estimates for three different tests. In addition the task had an additional constraint that only one reader could be purchased making it similar to a discrete choice. Fitting each of the test volumes separately would have resulted in different reader share predications. Instead we took a hybrid approach that has been commonly used in practice. We estimated a discrete choice model for the reader purchase and then weighted the resulting choice probabilities by a per-respondent weight that was computed as the maximum observed number of tests they purchased. As most respondents provided the same test quantities for all tasks where they made a reader purchase this seemed to be a reasonable estimate of quantity. It should be noted that if the number of tests purchased exhibited strong dependence on configuration of the reader or test prices then this would have been a less than optimal method. It would have been better to employ a two-step procedure similar to that described in Eagle (2010). In this case however, approximately 66% of the respondents showed no volume sensitivities to test prices or configuration beyond those measured in the discrete choice scenario.

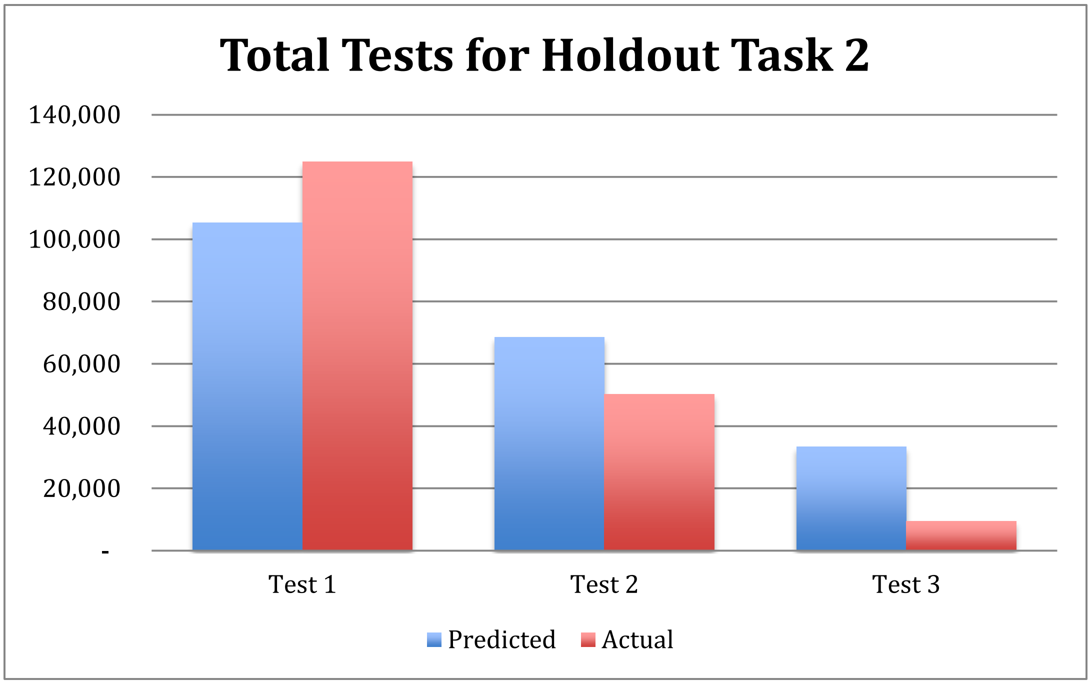


Figure 5. Volume Predictions for Holdout 2

The exact procedure is as follows. We first ran CBC/HB on the reader choice data. A number of runs were made including models with and without interactions. The model with the best fit was brought forward. Separately we calculated the maximum response for each respondent for each test type. This provided the weight for each respondent. (We also tried using the volume response as the weight, but the fit was significantly worse.) In Sawtooth Software's SMRT market simulator we ran a simulation for each holdout task using Randomized First Choice and saved the individual choice probabilities. (We also tested a first choice simulation, but first choice provided worse holdout prediction.) We then multiplied the individual choice probabilities by the volume weights to get the total volume prediction for each holdout task and each test. This provided a total of 9 volume predictions for each holdout task. The results for the share prediction are found in Table 2. Fixed Cost Conjoint Analysis provides better share predictions than CBC for both holdout tasks. The differences were rather surprising as this is fairly straightforward discrete choice exercise.

RMSE	FCCA	CBC
Holdout 1	**8.76 (5.63%)**	15.16 (10.31%)
Holdout 2	**5.22 (3.6%)**	7.29 (5.0%)

Table 2. Reader Share Predictions - FCCA vs. CBC

The summary root mean square error for the volume predictions is presented in Table 3. Fixed Cost Conjoint analysis is nearly twice as good for the first hold out task and slightly better for the second holdout choice task. If a reader was purchased, the volume for each test showed little variation and generally approached the maximum response. This represents an ideal dataset for CBC as the volume estimates were insensitive to the different product specifications. In an exercise where the volumes were more sensitive to the level of utility or prices we would expect that the FCCA would exhibit even more convincing results.

RMSE	FCCA	CBC
Holdout 1	**768.24**	1375.66
Holdout 2	**642.89**	687.27

Table 3. Test Volume Predictions for FCCA and CBC Comparison

EXAMPLE POLICY SIMULATION

In the introduction we highlighted the effect that fixed costs have on entry into a market place. In general fixed costs provide a barrier to consumers entering the marketplace. In this setting the test sales are significantly more attractive than the reader sales. The tests are consumable and provide a steady revenue stream that is renewable year after year. The reader sales on the other hand are a one-time sale that incurs significant support costs over the life of the reader. One general strategy that this firm could incur would be to subsidize the reader cost hoping to make up for the deficit on the repeated sale of tests. This is similar to the strategy employed in razors and video games consoles.

We can compare the revenue impact of the changing from the expected product configuration to a free reader strategy. Unfortunately we don’t have full cost information, so we can’t find a profit optimizing strategy.

The results from the simulation are presented in Table 4. As expected offering the reader for free dramatically increased the number of customers purchasing tests. The tests volumes also increased, but not as quickly as the increase in the number of customers. This implies that as the price of the reader decreases lower and lower volume customers begin using the product. Interestingly revenue only increased 49% with an over 3-fold increase in the number of customers. One of the reasons for this is that the reader contributed to a significant portion of the revenue in the base case. The real result however is that revenue per customer decreased 64%. With the cost of manufacturing the reader and the expected support costs, it is likely that the company would incur a net loss per customer instead of the profit predicted in the base case. It is likely that the base case does not represent the optimal strategy, but without cost information it is impossible to know what the optimal strategy would be.

In this situation fixed costs cause low volume customers to self-select out of the market while high volume customers purchase the product. This is especially useful information for a new company like this one where it allows them to focus on the most profitable customer while they have limited capacity.

	Base Case	$0 Reader	%Change
Customers	17.21	71.3	314%
Test 1	25,115	66,765	166%
Test 2	10,309	29,566	187%
Test 3	1,047	4,692	348%
Revenue	-	-	49%
Revenue/Customer	-	-	-64%

Table 4 - Simulation results of base case vs. $0 reader

This specific problem does not allow us to test the presence of local monopolies from fixed costs as all respondents face the same fixed costs. In a switching costs model, respondents face heterogeneous costs depending on past purchase decisions. Modeling these types of switching costs is a straightforward exercise if the data were available. In the simulator you would be able to see that customers that had previously purchased the fixed component were less likely to switch to a competing brand, as they would be required to incur the fixed cost again.

Conclusions

Volumetric Conjoint Analysis provides a theory grounded, straightforward way to estimate volumetric models. The economic theory it is founded on is especially suited to conjoint analysis and hierarchical Bayesian methods. Conjoint provides an easy a way to collect rich panel datasets at a relatively low cost. The panel structure of the data in turn allows us to fit individual level models using HB. This matches nicely with the economic theory of how individuals make decisions and how these decisions aggregate up to produce market outcomes. As these models develop, more complex and rich models will also develop.

We highlighted in this paper a method for modeling a common market situation that has so far been relatively neglected in both practice and in the academic literature. Models with mixed discrete/continuous demand highlight an interesting phenomenon that occurs in the real world, and cannot be accurately modeled with existing choice based techniques. The model presented here extends previous work on the modeling of volume-based choices by highlighting that consumers do not always face continuous prices. Bayesian data augmentation provides a means of solving these types of problems.

There are a number of areas where this work could be extended and improved upon. For practitioners of market research the most pressing need is for packaged software that can estimate these models. The models have shown themselves to be robust and useful in many applications. From an academic perspective there are a number of additional challenges and areas for future research. We are just touching the surface of the many different types of pricing and promotions strategies that exist in the real world. Many of these common situations have similarities with the current research in that the budget constraint is not perfectly continuous. A similar strategy could be employed to solve these problems.

References

Bell, David R., Teck H. Ho, and Christopher S. Tang (1998), "Determining Where to Shop: Fixed and Variable Costs of Shopping," *Journal of Marketing Research,* 35(3).

Chandukala, Sandeep R., Jaehwan Kim, Thomas Otter, Peter E. Rossi, and Greg M. Allenby (2007), "Choice Models in Marketing," in *Economic Assumptions, Challenges and Trends,* Now, 97–184.

Eagle, Thomas C. (2010), "Modeling Demand Using Simole Methods: Joint Discrete/Continuous Modeling," in *Sawtooth Software Conference October 2010*.

Gilbride, Timothy J., and Greg M. Allenby (2004), "A choice model with conjunctive, disjunctive, and compensatory screening rules," *Marketing Science*, 23(3), 391–406.

Satomura, Takuya, Jeff D. Brazell, and Greg M. Allenby (2011), "Choice Models for Budgeted Demand and Constrained Allocation," *Working Paper*, 1–31.

Tanner, Martin A., and Wing Hung Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82(398), 528–540.

A COMPARISON OF AUTO-RECOGNITION TECHNIQUES FOR TOPICS AND TONES

KURT A. PFLUGHOEFT
MARITZ RESEARCH
FELIX FLORY
EVOLVE24

INTRODUCTION

An explosion in consumer-generated media has created a great demand for text analytics. Open-ended comments often need to be summarized by topic and tone (i.e. sentiment). Historically, such efforts required much human intervention making it impractical to accomplish the task in a cost effective manner. More recently, several proprietary and open applications have been developed to expedite the process. In this research, a case study is used to explore and compare the results of text analytics with human coders.

Determining the topic and tone of comments has been a common type of market research for many years. To start this process, a list of topics is created by a quick perusal of the comments. Next, human coders would read the comments and assign them to one of many topics in that list. For example, hotel comments might be assigned to topics such as reservation, arrival and departure. Finally, reports were created to examine those topics such as their frequencies as shown in Figure 1.[41]

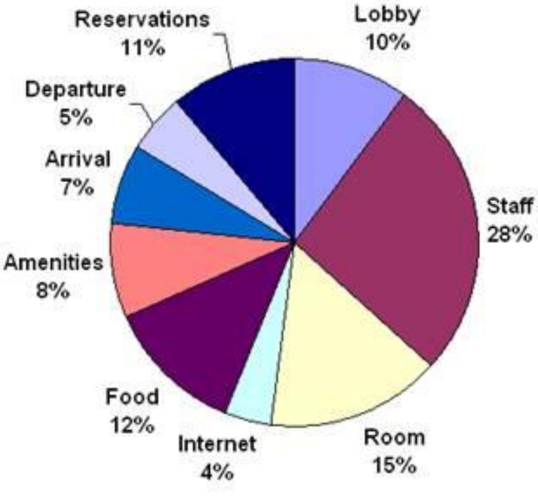


Figure 1: Hypothetical Distribution of Comments by Topic

Although these frequencies provide some information, this graph alone is not very actionable. For example, it is unknown whether a majority of topic comments were either

[41] Other charts may be used especially when there is more than one topic per comment.

positive or negative. One solution is to create more specific topics with an implied tone such as "bad staff service." However, this creates a greater burden for the human coder. Another solution is to adopt a proxy which may represent the tone of the comment. For example, if the comment was collected in an unrestricted manner, the overall satisfaction (OSAT) score for the survey may be used. The rationale is that the comment is capturing the salient customer experience issues that helped form their OSAT rating.

Figure 2 shows the top three box OSAT scores that are associated with the comments. Here, the top three box score for staff service is 81% whereas the top three box score for food is 46%. Thus, it is reasonable to assume that the staff service comments tend to be positive whereas the food-related comments tend to be negative.

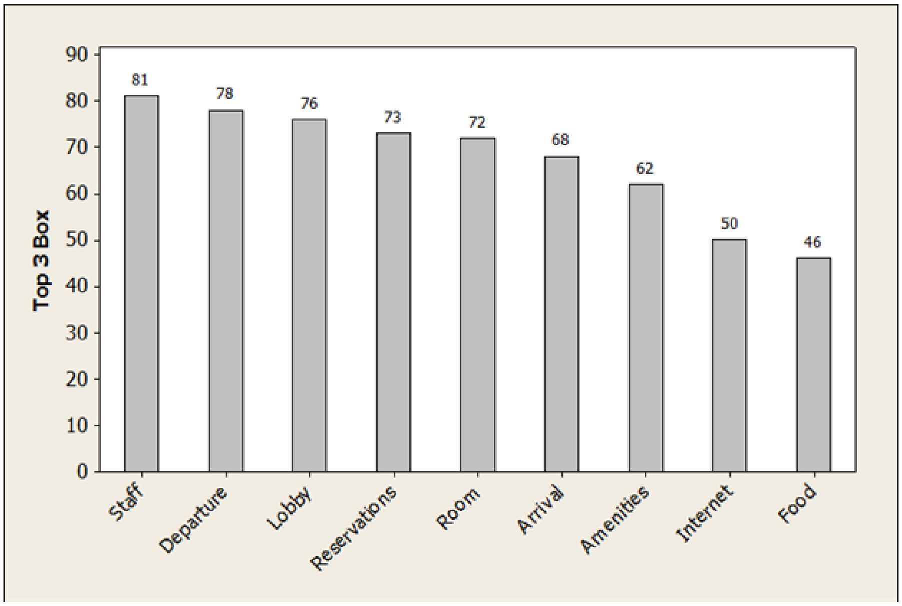


Figure 2. Hypothetical Top 3 Box Scores for Survey Comments Mentioning a Topic

Determining topic and tone usually requires at least two approaches as was done by coding the comments for topics and examining OSAT for tones. Likewise, the exploration of technologies can be done separately for tones and topics. Although a simultaneous evaluation is desirable, that complicates the discussion and comparison of techniques. Consequently, the first part of the paper will focus primarily on tone and the second part will discuss topic identification.

Inverse Regression

Three methods for toning comments are compared against human coders: Clarabridge, Lexalytics and Inverse Regression. The first two approaches are proprietary so no explanation of those approaches is given. On the other hand, the application of inverse regression for text analytics has been discussed in academic journals and implemented in the R package "textir."[42]

[42] See "Inverse Regression for Analysis of Sentiment in Text," http://arxiv.org/abs/1012.2098, M. Taddy, 2011.

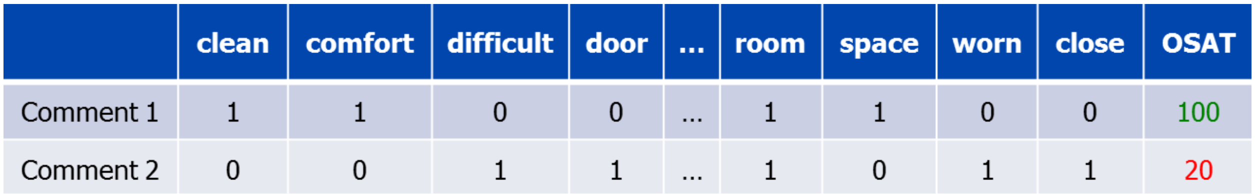
Since inverse regression is a statistical technique, it can be applied to many languages with minimal effort. A brief explanation of this approach is discussed below.

Inverse regression identifies the tones of keywords by their association with one or more proxy measures such as OSAT. The keywords must be identified beforehand by the use of a document term matrix. The keyword tones can be simply reported or they may be summarized to determine the overall tone of a sentence, the comment or even a topic.

The use of a document term matrix is quite common in text analytics and it quantifies each comment by the keywords present in all comments. Keywords must convey some meaning thus uninteresting words such as “a” and “the” are often ignored. Additionally, it may make sense to represent only one form of the keyword thus stemming can be applied but it is not required. For example, the word “write” may appear as “wrote”, “written” or “writing” but they can all be considered as a variation of the root word “write”.[43]

Figure 3 shows two example comments and their corresponding document term matrix - only a portion of the matrix is shown. The document term matrix records the number of times a particular keyword appears in a comment. Needless to say, the number of keyword columns grows very quickly as comments are added. Consequently, the document term matrix will become quite sparse, i.e. most of the column entries for a row will be zero indicating words which didn’t appear in the comment.

> Comment 1: “The hotel was very clean, comfortable and, rooms were spacious.” OSAT score: 100
>
> Comment 2: “The Room hardware was worn - doors wouldn't close, or were difficult to operate.” OSAT score: 20



	clean	comfort	difficult	door	...	room	space	worn	close	OSAT
Comment 1	1	1	0	0	...	1	1	0	0	100
Comment 2	0	0	1	1	...	1	0	1	1	20

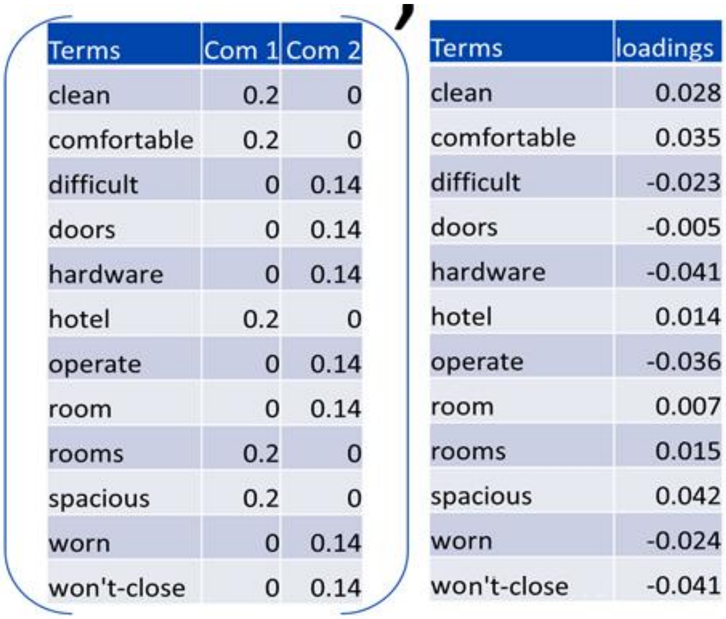
Figure 3: Document Term Matrix

The rightmost column, showing the OSAT score, is technically not part of the document term matrix but the score is appended to show that it is a required input for inverse regression. The score is needed only for training purposes; once the keyword tones are learned there isn’t a requirement to further use OSAT.

The next step of the inverse regression technique is to take this sparse matrix, or a reasonable subset[44], and learn the associations of the keywords with the OSAT scores. The inverse regression will result in a matrix of keyword tones (called loadings) where the sign of the loading indicates if it is a negative or positive keyword. Figure 4 shows an example of these loadings on the right with a transposed document term matrix of relative frequencies on the left.

[43] Actually, stemming may encounter some problems as the word “writing” could be a noun and not a conjugate of the verb write. Other text processing steps like a spell check may be applied before stemming.

[44] Usually it is sufficient to examine the subset of keywords that occur more frequently.



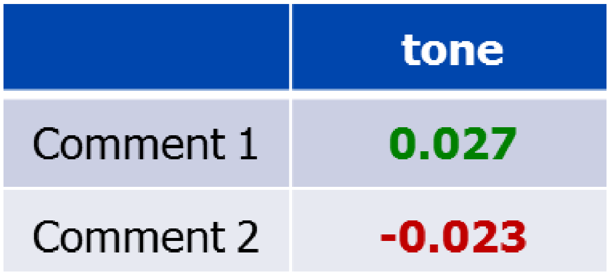
Terms	Com 1	Com 2
clean	0.2	0
comfortable	0.2	0
difficult	0	0.14
doors	0	0.14
hardware	0	0.14
hotel	0.2	0
operate	0	0.14
room	0	0.14
rooms	0.2	0
spacious	0.2	0
worn	0	0.14
won't-close	0	0.14

Terms	loadings
clean	0.028
comfortable	0.035
difficult	-0.023
doors	-0.005
hardware	-0.041
hotel	0.014
operate	-0.036
room	0.007
rooms	0.015
spacious	0.042
worn	-0.024
won't-close	-0.041

Figure 4. Relative Frequencies of Document Term Matrix and Inverse Regression Loadings

The loadings in the matrix provide valuable information. For example, the words “room” and “rooms” both have positive loadings; whereas “doors” and “hardware” have negative loadings. The latter issues could merit further investigation to determine the reason for their unfavorable loadings. Finally, some keywords take on an “expected” tone like the word “difficult” which has a negative loading.

The keyword loadings can be aggregated to determine the tone of the entire comment. First, the keyword loading matrix is multiplied by the relative frequency of each keyword in a comment. Second, these products are added together to result in the following comment tones as shown in Figure 5. The comment tones appear to be correctly identified as the first comment was clearly positive and the second comment was negative.



	tone
Comment 1	0.027
Comment 2	-0.023

Figure 5: Comment Tones

It should be noted that there are other ways to aggregate the keyword loadings within each comment. For example, there is a body of literature which indicates that people place disproportionately more emphasis on unfavorable information. Consequently, negative keywords might be weighed more heavily when determining the tone of the comment.[45]

TONE COMPARISONS

To determine the effectiveness of inverse regression, it is compared against two proprietary techniques: Clarabridge and Lexalytics as well as human coders.[46] One thousand hotel comments were chosen at random from the web. Results obtained from the human coders are assumed to be absolute truth in this analysis.[47]

Since the Clarabridge and Lexalytics solutions already exist and are proprietary, no additional training was conducted for these packages. However, the parameters for the inverse regression model did not exist, thus 14,000 separate hotel comments and their OSAT ratings were used for training purposes.

The tone ratings obtained by Clarabridge, Lexalytics and human coders are binned into three categories: positive, negative and neutral. The results from the inverse regression technique were binned into their natural levels of positive and negative. The distribution of comment tones, as assigned by human coders, is: 20.6% negative, 5.9% neutral and 73.5% positive.

One way to compare the three auto-coding techniques to the human coders is to look at their accuracy, i.e. what is the combined percentage of times that two approaches are positive? negative? and neutral? The accuracy measures from the experiment are 76.1%, 77.7% and 81.7% for Clarabridge, Lexalytics and inverse regression respectively. All accuracy measures are good and fairly close to each other but the "lift" is relatively small when considering 73.5% of all comments are positive. (It should be noted that the inverse regression technique is given a slight advantage due to the omission of a neutral bin for this test data.)

Accuracy is one reasonable measure but it does have its flaws. Consider a naive forecast which would classify all comments as positive and have an accuracy of 73.5%. Sounds pretty good; however, it makes a pretty severe error 20.6% of the time. Negative comments are actually classified as if they were positive. Invariably, those errors and the opposite situation occur under many techniques but hopefully not to that extent.

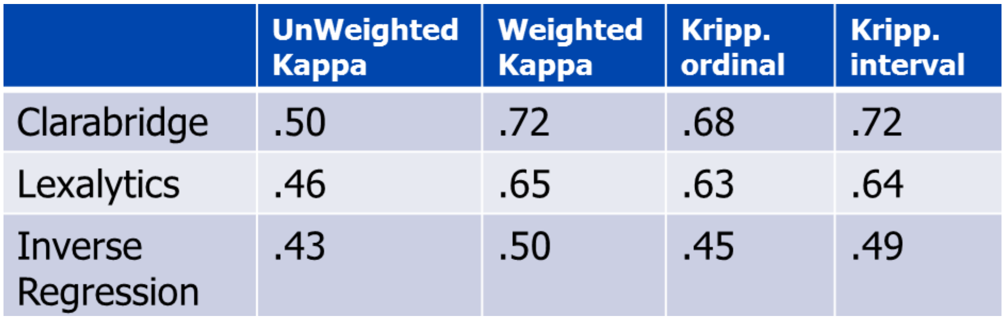
To gain a more holistic comparison, other measures exist to discount random agreement and penalize the severity of the error. These measures include several variations of Cohen's Kappa and Krippendorff's Alpha. Both types of measures usually range from 0 for no agreement to 1 for perfect agreement.

The agreement measures for the 1000 test cases are show in Figure 6.

[45] See "An Attribution Explanation of the Disproportionate Influence of Unfavorable Information" by Miserski, 1982 or literature concerning Negative Word of Mouth (NWOM).

[46] Disclaimer: The purpose of this research is to do a comparison of text analytic techniques for a case study. Results achieved here are dependent upon the version of software used, processing decisions/assumptions made by the authors and the nature of the particular comments. The authors make no claims, either positive or negative, about the overall performance of any of these software solutions or techniques beyond this case study.

[47] Coding was outsourced to a firm which specializes in this task.



	UnWeighted Kappa	Weighted Kappa	Kripp. ordinal	Kripp. interval
Clarabridge	.50	.72	.68	.72
Lexalytics	.46	.65	.63	.64
Inverse Regression	.43	.50	.45	.49

Figure 6. Inter-rater Agreement Measures

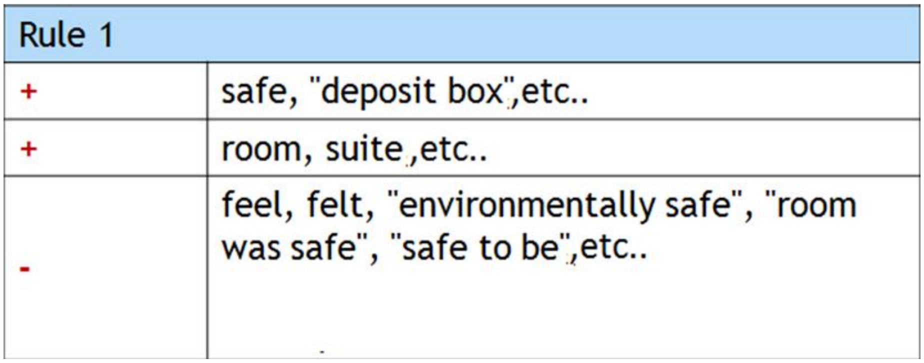
These inter-rater agreement measures show that Clarabridge performs the best followed by Lexalytics and inverse regression. This is partly due to the occurrence of the most severe type of classification error which is 3.4% for Clarabridge, 6.1% for Lexalytics and 12.4% for Inverse regression. Here, the use of only two bins for inverse regression puts it at a disadvantage. [48]

The above measures indicate moderate to substantial levels of agreement with the human coders. For a reference point, a naïve[49] forecast has Kappa values close to zero and has Krippendorff's values below zero.

Topic Identification

The last part of this research investigated the automatic classification and identification of topics. The former feature is already built into the Clarabridge application via an extensive, domain-specific rule base. Informally it can be reported that Clarabridge does a pretty good job when compared to human coders. Lexalytics can also identify topics but that feature has not been used in this research.

An example of a Clarabridge rule to identify a comment concerning a "physical safe" is shown in Figure 7. For this rule to fire, at least one word or phrase must appear in the comment from each of the rows prefaced by a plus sign and no word or phrase must be present from the last row.



Rule 1	
+	safe, "deposit box",etc..
+	room, suite,etc..
-	feel, felt, "environmentally safe", "room was safe", "safe to be",etc..

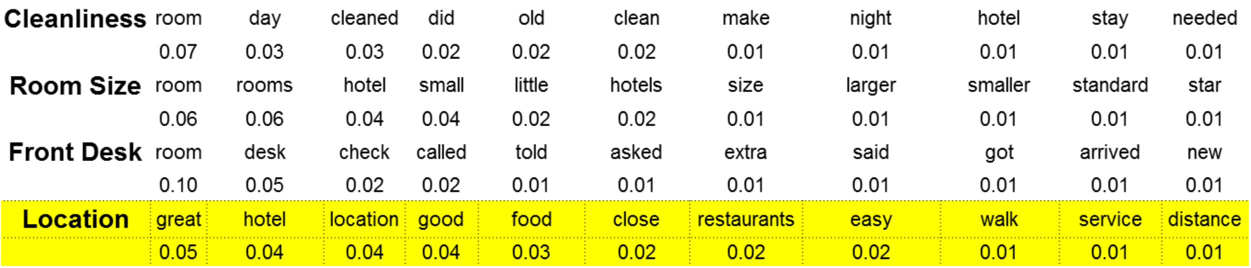
Figure 7: Portion of a Rule to Identify the Topic "Room Safe"

[48] Determining a neutral category for Inverse Regression takes a bit more research. Using marginal probabilities for the positive, negative and neutral categories from the training set can improve the results to be similar to those attained by Lexalytics.

[49] The naïve prediction was adjusted to include one negative and one neutral prediction as inter-agreement measures require entries under all tone levels.

As you can see that a rule-based approach takes significant effort to build for a particular industry – there are potentially hundreds of rules.   Statistical approaches require much less domain knowledge and can be trained to identify and classify comments – though they may not be topics that match your predefined list. One popular statistical approach is topic modeling which uses Latent Dirichlet Allocation (LDA). You can broadly think of this approach as a type of factor analysis for words.[50] The number of topics must be identified in advance by the researchers and the LDA results will indicate the probability that each word is associated with a particular topic.

Just like factor analysis, it is the responsibility of the analyst to assign a meaningful name to the topic. A portion of the topic model output is shown in Figure 8. The yellow rows indicate a topic that was not previously used by the analysts.



Cleanliness	room	day	cleaned	did	old	clean	make	night	hotel	stay	needed
	0.07	0.03	0.03	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01
Room Size	room	rooms	hotel	small	little	hotels	size	larger	smaller	standard	star
	0.06	0.06	0.04	0.04	0.02	0.02	0.01	0.01	0.01	0.01	0.01
Front Desk	room	desk	check	called	told	asked	extra	said	got	arrived	new
	0.10	0.05	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Location	great	hotel	location	good	food	close	restaurants	easy	walk	service	distance
	0.05	0.04	0.04	0.04	0.03	0.02	0.02	0.02	0.01	0.01	0.01

Figure 8: Topic Modeling Output

The new topic was coined “location” and its meaning can be ascertained by the probability of its associated keywords.   Here, “location” implies that it is a close distance or an easy walk to other services such as restaurants. Thus, topic modeling can be extremely useful not only in coming up with a predefined list of topics but also in the identification of emerging topics.

Conclusion

There have been many advances in text analytics to help identify the topic and tone of comments. Results appear to be promising and in many cases the tones assigned by automated techniques show substantial levels of agreement with human coders. Identification of topics is more difficult but it can be accomplished by either the formation of rules or the use of other techniques such as statistics. All topic approaches requires analyst time for setup but a rule-based/dictionary approach requires more effort and experimentation.

The level of agreement of topics to human coders is not included in this paper. Future research efforts to explore that aspect and fine tune techniques when possible are welcomed.

[50] Technically, LDA can be considered a type of principal components analysis; see D. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, to appear.