PROCEEDINGS OF THE SAWTOOTH SOFTWARE CONFERENCE

October 2010

FOREWORD

These proceedings are a written report of the fifteenth Sawtooth Software Conference, held in Newport Beach, California, October 6-8, 2010. This conference was held in concert with the third Conjoint Analysis in Healthcare Conference, chaired by John Bridges of Johns Hopkins University. Conference sessions were held at the same venue, and ran concurrently. The presentations of the healthcare conference are published in a special edition of The Patient—Patient Centered Outcomes Research.

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 menu and bundling choice tasks, choice/conjoint analysis, MaxDiff, and hierarchical Bayes estimation.

The papers are in the words of the authors, with generally very little copy editing done on our part. We are grateful to the authors who sacrificed time and effort in making this conference one of the most useful and practical quantitative methods conferences in the industry. While preparing this volume takes significant effort, we'll be able to review and enjoy the results for years to come.

Sawtooth Software

March, 2011

Contents

SUMMARY OF FINDINGS

The fifteenth Sawtooth Software Conference was held in Newport Beach, California, October 6-8, 2010. The summaries below capture some of the main points of the presentations and provide a quick overview of the articles available within the 2010 Sawtooth Software Conference Proceedings.

What Drives Me? A Novel Application of the Conjoint Adaptive Ranking Database System to Individual Vehicle Consideration Set Formation (Ely Dahan, UCLA, Princeton University): Most conjoint analysis studies are commissioned to help managers understand the collective behavior of segments and groups of respondents. Ely demonstrated an application over the web for using conjoint analysis to help individual buyers select vehicles in real-time. He reported that the key to using conjoint analysis as an online recommendation agent is to make the interview very short (5 to 10 clicks), interesting, and to give respondents quick and rewarding feedback. Ely's approach used a combination of adaptive questioning, prior population knowledge (including real sales data), and engaging graphics. At the end of the short survey, the recommendation agent gives participants a prioritization of vehicles that have a close match with the respondent's preferences. Ely's team tested multiple versions of the questionnaire to improve the layout of the information, and Ely showed the evolution of the questionnaire appearance.

The Success of Choice-Based Conjoint Designs among Respondents Making Lexicographic Choices (Keith Chrzan, John Zepp, and Joseph White, Maritz Research): Keith and his coauthors described how minimal overlap design strategies assuming a compensatory MNL choice process can do a poor job of capturing the preferences of respondents making non-compensatory choices. They first demonstrated this for artificial respondents and then assessed the magnitude of the potential problem for a variety of design strategies using two large commercial studies in which pretest respondents exhibited non-compensatory decision processes. When holdout choice tasks involved no level overlap, minimal overlap designs did well in predicting them. But, designs with overlap also did fairly well. However, when holdout tasks included level overlap, the designs featuring overlap did better. The authors pointed out the significance of this finding in light of the fact that most real-world market simulations involve a healthy amount of level overlap. Keith and his coauthors argued that it wasn't enough just to provide overlap on all attributes, but that qualitative work and pretesting could identify the attributes that respondents most likely screen on. Then, designs should feature the most overlap on attributes that involve the most screening. As a sidelight, the authors showed that asking both best and worst choices within CBC tasks may improve predictive accuracy of holdout choices over asking first choices alone.

Menu-Based Choice Modeling Using Traditional Tools (Bryan Orme, Sawtooth Software, Inc.) Bryan described how researchers today are increasingly being asked to model choice processes that involve menu-based selection. In menus, respondents make between one to multiple choices to configure their preferred choice. Bryan explained that randomized designs make these relatively easy to design, but data processing and analysis can be challenging. Using a real dataset with 1600 respondents, he investigated different ways to analyze and model choices for options on new automobiles using Sawtooth Software tools, including counting analysis, logit, latent class, and CBC/HB. A dataset with 1600 respondents was used (800 calibration respondents, and 800 holdout respondents). Three different model formulations were

tested ("Volumetric CBC," Serial Cross-Effects, and Exhaustive Alternatives). Bryan demonstrated how to code all three models for estimation using MNL. All methods did quite well in predicting holdout choices, and there wasn't much difference in the predictive accuracy of both aggregate and individual-level parameter estimation. Bryan explained under which data conditions different model formulations might be preferred.

Analysing Pick n' Mix Menus via Choice Analysis to Optimise the Client Portfolio (Chris Moore, GfK NOP): Chris laid the groundwork for his presentation by stating that menu-based choice research is becoming more prevalent in our industry, following the trends toward mass customization. He then described a case study conducted by his firm for TGI Friday's restaurants. The study design followed a two-stage process (both stages within the same questionnaire). In the first stage, a CBC study was used to predict which of several restaurants (including TGI Friday's) respondents would choose to visit, given changes in the main aspects of the menu. In the second stage, a menu-based choice study was used to predict what buyers would pick from the TGI Friday's menu, given price changes just for TGI Friday's items. A series of cross-effects logit models were employed in the second stage to optimize pricing for key dishes to ultimately increase net profit for the client. The model predicted holdout scenarios well, with an R-squared better than 0.95. Based on the recommendations of the model, TGI Friday's implemented an optimized menu within a test store. After six months of sales data, the test store showed 15% higher profit than the control stores. More recently, Chris re-analyzed the data and pruned 45 of the 80 cross effects that were not found to have a significant effect on respondent choices. The pruned cross-effects model showed slightly better prediction accuracy for holdouts.

An Empirical Test of Bundling Techniques for Choice Modeling (Jack Horne, Silvo Lenart, Bob Rayner, and Paul Donagher, Market Strategies International): Jack and his coauthors compared two methodologies for product/service bundling research: a CBC approach that uses availability effects to estimate a feature's underlying intrinsic utility (conditional on prices offered); and the binomial choice-modeling approach for the analysis of build-your-own data. They found that the projected volume of items picked from the menu is very different if respondents choose among pre-specified bundles versus having the opportunity to build their choices a la carte. Their data seemed to confirm the theory behind bundling strategies: creating bundles for buyers to consider rather than letting buyers customize their purchase a la carte can help overcome individual reservation prices on items and lead to overall greater volume of sales for the items, with greater profitability to the firm, so long as they can only choose individual items from a single brand. When buyers can choose items a la carte across several brands, as they would in most marketplaces, profitability for the individual firm can be similar, if not slightly larger, to that obtained from offering fixed bundles.

Anchoring Maximum Difference Scaling Against a Threshold – Dual Response and Direct Binary Responses (Kevin Lattery, Maritz Research): Maximum Difference Scaling is widely used to measure the relative values of items/attributes. Despite the strengths of MaxDiff, some analysts would prefer data that represented more than just relative scores; they would prefer absolute scores scaled with respect to each respondent's importance threshold. In Kevin's presentation, he tested two methods for anchoring MaxDiff scores to a threshold: Dual-Response MaxDiff suggested by Louviere and a more direct method asking respondents to choose which attributes are above a threshold (using 2-point scale grid questions). He found that the original way for coding the Louviere approach as described in a Sawtooth Software white paper was incomplete and performed poorly for purposes of HB estimation, but that a very recent update by

Sawtooth Software to the suggested coding procedure produced much better results. He determined that theoretically (using synthetic respondent data) the direct method would be superior, especially as the number of attributes shown in a MaxDiff task increases. With six or more attributes shown per screen the indirect dual-response method should not be used, and even five attributes per screen may not capture individual anchoring that well. In comparing the two methods with human respondents, showing only four attributes per screen, results were very similar. The rank order of utilities at the respondent level was nearly identical. However, the anchoring in the direct method was more biased by the context of the total set of attributes. So if it is important for one to have a more neutral anchor for utilities then the indirect dual-response method may be slightly better, assuming four (and certainly no more than five) attributes are shown per screen.

Directing Product Improvements from Consumer Sensory Evaluations (Karen Buros, Radius Global Market Research): Using consumer evaluations to guide product development is problematic when a product fails to achieve its goals. Karen described an alternative to Penalty Analysis—sensory drivers and simulation—to understand which attributes play a greater role in driving satisfaction. A key problem with penalty analysis is that it allows changes in perception of product features where the modifications are in fact contrary to respondents' held beliefs about the relationships among the product characteristics. For example, penalty analysis could permit the analyst to simulate the change to product desirability due to simultaneously increasing both its perception of sweetness and tartness. Such a product change would seem impossible to achieve. Rather, she recommended an approach where product perceptions on attributes were first submitted to factor analysis to obtain a reduced set of factors that captured the underlying correlation structure among individual attributes. Then, the resulting simulator allowed the analyst to make changes to the overall factors rather than the original attributes. This approach more correctly avoided impossible formulation modifications, keeping the product recommendations better in line with the relationship of variables as perceived by respondents.

A Study of the Diffusion of Alternative Fuel Vehicles: An Agent-Based Modeling Approach (Rosanna Garcia, Northeastern University and Ting Zhang, Xi'an Jiaotong University): Rosanna and Ting Zhang advocated the use of agent-based modeling as an extension and enhancement of the traditional simulations conducted with conjoint analysis data. The topic for their modeling was the diffusion of eco-innovations, such as green automobiles. Their model involved the interaction among multiple agents: buyers, car manufacturers, and government agents who enact policies and regulations. The conjoint data was used to provide the model with an empirical foundation. They used a software program called NetLogo to run a series of agent-based model simulations, where different mechanisms were considered for speeding the adoption of alternative fuel vehicles. Additionally, they compared the results of agent-based modeling to that of a recently published game theoretic model, finding excellent agreement. The agent based simulation provided both a micro and macro analysis of the factors influencing the diffusion of the eco-innovation.

The Impact of Respondents' Physical Interaction with the Product on Adaptive Choice Results (Robert J. Goodwin, Lifetime Products, Inc.): Bob conducts research at Lifetime Products, a manufacturer of products such as folding furniture (chairs and tables). He reported on two studies designed to see whether there were large differences in conjoint data collected using in-person mall intercept data (where respondents could physically interact with the products) versus online panels. Since in-person interviewing is much more expensive and time consuming

to conduct, if online results did at least as well as in-person, it would make sense for Lifetime to use online panels instead. Two split-sample ACBC studies were conducted using online and mall-intercept field methods. Market simulation results were then validated using actual product sales and market share distributions. Bob found a few significant differences in the utilities between the two data collection modalities. In general the online panel data seemed to predict actual sales a bit better than in-person. But, there was instance in which a new kind of mesh chair seemed to require more tactile interaction with the product to understand its characteristics. The online approach didn't seem to capture the strength of preference for the new mesh chair, because it just didn't sound very appealing when described in an online interview. Bob concluded that panel data offered a good, cost-effective alternative, but there might be instances in which in-person interviewing was a preferred option for his firm.

Using Eye Tracking and Mouselab to Examine How Respondents Process Information in Choice-based Conjoint Analysis (Martin Meißner, Sören W. Scholz, and Reinhold Decker, Bielefeld University, Germany): How respondents actually process information in CBC interviews is somewhat difficult to assess by only observing their choices. Martin and his coauthors used eye tracking technology to examine what respondents looked at, and the patterns of gazes as they answered CBC questionnaires. They also employed a second methodology called Mouselab to try to record the way respondents evaluated complex CBC tasks. They investigated a number of issues, such as the degree to which respondents appear to be employing non-compensatory decision heuristics, and whether respondents shift their attention to different attributes as the interview progresses. And, especially, they wanted to know whether including information from eye-tracking could actually improve the conjoint utilities. Their results confirmed earlier findings that respondents pay relatively more attention to brand in the first tasks, but relatively more to price in the later tasks. They also found a slight but statistically-significant inclination toward simplification in later tasks (respondents paying attention to fewer attributes).

The Value of Conjoint Analysis in Health Care for the Individual Patient (Liana Fraenkel, Yale University School of Medicine, VA Connecticut Healthcare System): As patients play a greater role in health-related decisions, their assessment of risks and benefits of treatment and cancer screening becomes a critical factor. In this presentation, Liana described the value of conjoint analysis as a tool to enable patients to understand better their own preferences and to understand more completely the impact of specific attributes on choices. Liana demonstrated that asking patients to complete a conjoint survey leads to patients feeling more informed, satisfied, and more prepared to discuss their treatment with their doctor. If doctors can understand patient preferences better (by viewing a report showing respondent preferences computed via conjoint analysis) and can incorporate patient preferences into the prescribed treatment, then the likelihood of compliance with the treatment will likely increase. Liana also described how she has tried different modifications to the ACA survey to try to improve the patient experience and improve the data. Particularly, she has tried different modifications to the Importance question. One modification that has worked well is to ask respondents first to identify the one most important attribute. Then, she shows a grid question with all attributes listed, and asks respondents to assign that one most important attribute a 10, and to rate the others with respect to it. She hopes to employ MaxDiff as well in future research, and thinks that this may be even easier for patient populations to manage.

Personalizing Treatment for Depression: Developing Values Markers (Marsha Wittink, University of Rochester Medical Center, Knashawn Morales, and Mark Cary, University of Pennsylvania School of Medicine): Marsha explained that tailoring the features of treatment to the way patients experience and think about depression may help to increase acceptance of and adherence to treatment. Through conjoint analysis, one might identify which attributes of treatments are most important, and this could help doctors design better interventions. Despite existing effective depression treatments, poor adherence is prevalent, she explained. Analogous to genetic markers (profiles of genetic variation related to treatment response) her research proposes to identify values markers—profiles of values related to the attributes of treatment that patients value most. Values markers can provide specific guideposts for how to design personalized depression treatment. Specifically, she described efforts using CBC and empirical Bayes methods (using SAS Glimmix). She also developed a three-group solution using latent profile analysis. The next steps for her research will involve tailoring interventions for patients based on preferences learned from CBC interviews, and to see if this process improves treatment over using standard demographics variables for guiding preferred treatments.

Conjoint Design Effect on Respondent Engagement throughout a Survey (Paul Johnson, Western Wats): In this presentation, Paul mentioned that even though past research has shown Adaptive CBC to be more enjoyable for respondents than CBC, enjoyment doesn't always mean respondent fatigue is reduced. Enjoyable, engaging tasks can actually be more fatiguing. He conducted a split-sample experiment wherein respondents received either an ACBC or a CBC design with the conjoint either appearing before or after key response variables such as purchase intent. He examined fatigue metrics and other questions typically asked in a survey outside the conjoint portion to compare conjoint design and placement effects. He did find a few statistically significant effects. As has been shown before, respondents spent significantly more time in the ACBC survey than the typical CBC survey. And, for respondents who got the ACBC survey first, the likelihood of following simple instructions in later questions decreased, suggesting greater fatigue. He also included holdout choice tasks in the survey, including standard (non-adaptive) holdouts, and a "winners" holdout that was comprised of winning concepts from the previous three non-adaptive holdouts. Hit rates were very similar for CBC and ACBC in predicting the more challenging "winner" holdout. But, ACBC seemed to perform better in share prediction of the winner holdout. Paul concluded that there is no free lunch. If you use ACBC to get more out of respondents and engage them in a more thoughtful and enjoyable process, you may use up respondent's energy and responses to later questions may be negatively affected. Despite that disadvantage, Paul suggested that it is important to the industry to provide more enjoyable surveys.

Sales Promotion in Conjoint Analysis (Eline van der Gaast and Marco Hoogerbrugge, SKIM): Eline and Marco co-presented this paper, focusing on sales promotion as an attribute in conjoint studies. They started with a general theoretical background, followed by discussing various types of sales promotions: price discount, extra volume, etc. Promotions may involve a direct financial gain, and/or indirect benefits. The promotion generates extra attention for the product and the feeling of saving money. Eline and Marco showed results from a simple experiment in a survey that if one does a promotion that has the same financial savings to respondents as lowering the normal price, the effect of the promotion is much higher than simply reducing the price. They cited other research at their company that has found that promotions that specifically state the financial gain (either in % or rebate in currency terms) are more

successful in driving choice as compared to other promotion types with a direct gain e.g. showing (gross and) net price. They also warned that promotions provide a short-term benefit followed by a post-promotion dip. Even though promotions are difficult to study, conjoint analysis is effective in helping understand which promotion is more effective and which consumers you will attract with the promotion. Eline and Marco discussed that in the future, perhaps time elements could be incorporated into conjoint studies, to simulate more accurately purchase cycles and long-term effects of promotions.

How Many Questions Should You Ask in CBC Studies? – Revisited Again (Jane Tang and Andrew Grenville, Vision Critical): Jane recounted that two previous two papers of the same title concluded that respondents can reasonably handle a large number of choice tasks (up to 20), and there is diminishing return in the accuracy in predicting holdout tasks at between the 10-15 tasks mark. Jane and her coauthor reasoned this may be partly a result of respondents increasingly relying on heuristic simplifying rules. They designed a split-sample, three-cell CBC experiment including the following types of interviews: 6 tasks with 3 concepts each, 15 tasks with 3 concepts each, and 15 tasks with 5 concepts each. They conclude that longer CBC questionnaires will lead to slightly improved predictions of holdouts, but the later tasks have less sensitivity and more reversals in the price attribute. They did not find improvement for their study when using 5 concepts versus 3 concepts per task. Dual-response None usage also increased in later tasks, as has been found before for standard None implementation. Jane also reported that respondents tended to use simplification behaviors more in later tasks. In general, she and her coauthors recommend that researchers work to keep respondents happy and engaged, which means favoring shorter CBC surveys. Shorter CBC surveys can in part be compensated for by increasing the total sample size.

The Strategic Importance of Accuracy in Conjoint Design (Matthew Selove, USC Marshall and John Hauser, MIT Sloan): Matt's presentation focused on the notion that improved accuracy in conjoint analysis has important strategic implications. Uncertainty in part-worths makes product differentiation seem to be a more attractive option for competing firms to capture greater profits.

On the other hand, randomness in individual customer behavior makes product differentiation seem less attractive. Matt illustrated this outcome using a small dataset collected among university students dealing with hooded sweatshirts, fleece vests, and track jackets. Respondents received two different CBC exercises: a well-designed CBC study, and a poorly designed CBC survey that led to lower accuracy of estimates of preferences. The poorly designed questionnaire resulted in lower scale for logit effects. A market simulator was constructed to predict optimal choices by firms seeking to maximize profits. The data demonstrated that more random utilities leads to making differentiation in the product line less attractive. However, uncertainty in the utilities made differentiation seem more attractive. As a result, Matt argued that it is important to accurately estimate the true randomness in customer's purchase behavior (the logit scale parameter).

Product Portfolio Evaluation Using Choice Modeling and Genetic Algorithms (Christopher N. Chapman, Microsoft and James L. Alford, Blink Interactive) Chris and James described how they are using CBC and ACBC data together with genetic algorithm (GA) optimization to find near-optimal product portfolios in the presence of competition. Chris described a situation at Microsoft in which around 20 products were offered in a product line.

Managers wanted to know if they could pare back the number of products without much loss in sales in that category. Using the optimization routines and conjoint data, the authors found that there was sharply diminishing return for offering more than about six products in that line. Additionally, the authors were able to examine whether there were opportunities for new products in the portfolio to fulfill unmet needs. Chris and James used R code to run the optimization simulations, and said that the code was freely available to others who would like to use it. The authors also compared optimization results based on CBC and ACBC data. They generally found very strong congruence, but perhaps a slight edge in favor of ACBC in terms of respondent engagement, consistency of results across repeated simulations (from different starting points), and ability to obtain stable solutions with smaller sample sizes.

The Impact of Covariates on HB Estimates (Keith Sentis and Valerie Geller, Pathfinder Strategies): At previous Sawtooth Software conferences, researchers tested whether first segmenting the data and running HB within segments was better than using all respondents together in a single HB run (Sentis & Li 2001; Frazier et al. 2009). This year, Keith and Valerie revisited the issue in light of the new capability within CBC/HB v5 software to use covariates in the estimation process. They focused on 5 commercial datasets from studies involving both services and FMCG with sample sizes ranging from 420 to 5,502. They compared the predictive validity of holdout tasks when leveraging covariates and the new capability in the HB software versus default HB estimation assuming a single normally-distributed population. They tested covariates one at a time, for each dataset. The covariates involved demographics, category behavior, and attitudinal variables. They consistently found essentially no benefit for using covariates in terms of predictive validity of holdouts. They did, however, find that the difference in utilities between respondent segments was enhanced by the use of covariates. However, further analysis called into question how that was being accomplished. Keith and Valerie randomly scrambled the values for covariates, and found that even when using random segment assignment, the use of the covariate increased the discrimination on utilities between segments. This result suggests overfitting and further analyses showed that this overfitting increases as the sample size decreases. Thus, the use of covariates leads to greater differences among the segments, but there may be overfitting and some of the enhancement may be spurious.

Added Value through Covariates in HB Modeling? (Peter Kurz, TNS Infratest Forschung GmbH and Stefan Binner, bms marketing research + strategy): Peter and Stefan re-analyzed ten CBC data sets, comparing the use of covariates in HB to default HB assuming a single multivariate-normal distributed population. With covariates (segmentation variables), respondents are influenced (shrunk) toward preferences of respondents with similar characteristics rather than to a global population mean. In theory, this would seem to be more appropriate. Peter and Stefan looked to see whether in practice the use of covariates offered gains in terms of predictive validity of holdout choice tasks for CBC studies. For most all the datasets, there were no gains in predictive validity when including covariates, whether the covariates were demographics, benefits segments, or segments based on past behavior or purchase intention. They also re-analyzed data sets by systematically throwing away portions of the respondents and portions of the tasks. Reducing the amount of data in either way did not affect the outcome: covariates still did not improve predictive validity for holdouts relative to the default HB approach. Peter and Stefan also examined whether covariates could help improve matters when using proportional sampling with small segments of the population. They didn’t find that

covariates completely resolved previously identified problems with oversampling small segments, though there seemed to be improved results.

Modeling Demand Using Simple Methods: Joint Discrete/Continuous Modeling (Thomas C. Eagle, Eagle Analytics of California, Inc.): Tom described the joint discrete/continuous volumetric model for choice experiments and compared it to two other forms of simple volumetric models: regression-based and choice based models using several actual data sets. The choice modeling approach has the benefit of being a single-stage estimation approach, and can be accomplished rather easily with CBC/HB software alone. For four relatively simple volumetric CBC datasets, the choice-based models worked the best. But, for a more complex dataset, the joint discrete/continuous volumetric model performed better. The joint discrete/continuous model is estimated in two steps (e.g. CBC/HB followed by HB-Regression; though it also may be done with any MNL approach in the first stage combined with regression in the second stage). First, the volumes are normalized per task and treated as constant-sum chip allocation to estimate a set of logit weights. In the second stage, the logit weights are used as predictors of volume. Regression-based models were deemed inferior, and Tom recommended the other two approaches.

A Head-to-Head Comparison of the Traditional (Top-Down) Approach to Choice Modeling with a Proposed Bottom-Up Approach (Don Marshall, TVG Market Research and Consulting, Siu-Shing Chan, University of Pennsylvania and Joseph Curry, Sawtooth Technologies): At the 2009 Sawtooth Software Conference Professor Jordan Louviere presented some research indicating problems with the traditional approach to choice modeling leading to what he characterized as misleading results. He then presented a technique to eliminate these biases (a Bottom-Up approach that asked more than first choice for each task and used purely individual-level estimation). Don and his co-authors designed two studies to test Louviere's assertions, and to see if the standard first-choice CBC approach with HB estimation (Top-Down) produced erroneous results compared to Louviere's new method (Bottom-Up). The first study involved pizzas (6 attributes) and the second digital cameras (9 attributes). The results showed that Bottom-Up questionnaires were much longer than standard first-choice CBC questionnaires, leading to more dropouts in the Bottom-Up questionnaire. The predictive abilities of the two approaches were equal for the pizza study, and slightly favored the Top-Down approach for the camera study. When the Bottom-Up data were re-analyzed using HB, the results were directionally superior to the purely individual-level estimation approach. The authors concluded that Bottom-Up seemed to offer no benefit over traditional CBC with HB estimation, and was a longer and less satisfying experience for respondents.

*** HB-CBC, HB-Best-Worst-CBC or No HB at All?** (Ralph Wirth, GfK Marketing Sciences): Ralph compared different approaches for estimating part-worth utilities from CBC data, specifically HB vs. a purely individual-level method suggested by Louviere et al. The CBC data involved choices of best and worst concepts within each task (Best-Worst CBC). Key issues he sought to resolve were (1) whether HB choice models work well even under very challenging and sparse data conditions, (2) whether HB choice algorithms have systematic troubles (as Louviere et al. asserted) when individual-level errors (scale factors) differ a lot across the sample and (3) whether HB-Best-Worst CBC has an advantage over traditional ("best only") HB-CBC. Ralph generated a very large number of synthetic data sets, wherein he varied the error variance, the number of choice tasks, the number of attributes and levels, the sample size, and preference heterogeneity. He then compared for each condition how well HB-CBC, HB-Best-Worst-CBC

and the purely individual-level estimation method could recover true utilities and importances, as well as hit rates and share prediction accuracy of holdout tasks. He generally found both HB approaches to be superior to the purely individual-level method, and found no systematic problems for HB when the individual error variances were not constant across the sample. He also concluded that worst information in addition to best information from each task significantly improved results. Ralph also examined four datasets involving real respondents, finding in each case that the Best-Worst CBC did better (in predicting first-choice holdout tasks) than when analyzing the best-only information.

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

WHAT DRIVES ME? A NOVEL APPLICATION OF THE CONJOINT ADAPTIVE RANKING DATABASE SYSTEM TO VEHICLE CONSIDERATION SET FORMATION

ELY DAHAN
UCLA, PRINCETON UNIVERSITY

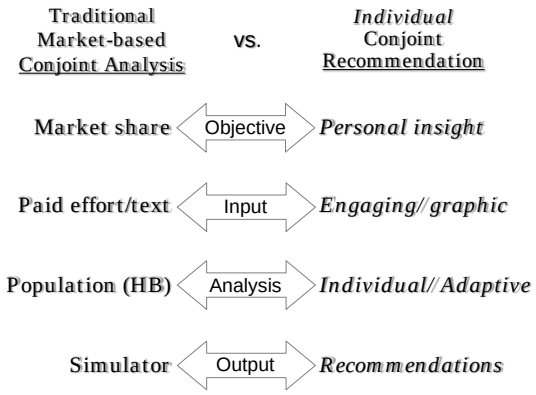
EXECUTIVE SUMMARY:

A new method of individual, adaptive choice-based conjoint analysis for vehicles on the web points to a future of highly efficient questioning with a new purpose: helping customers understand *themselves*. Several discoveries underpin the effectiveness of this approach: (1) the development of adaptive choice-based conjoint based on a predefined database of utilities, (2) development of a random utility generator that acts as a simulator of the actual market, including the ability to reproduce real market shares, (3) the use of actual products as conjoint stimuli, and (4) fine-tuning the tradeoff between allowing for respondent error versus enforcing consistent answers using computer assistance.

INTRODUCTION:

We propose a new method of real-time, web-based adaptive conjoint analysis for the purpose of assisting users in identifying a narrow set of vehicles for consideration in the early part of the purchase process. A primary goal is identifying ideal vehicles without resorting to purely non-compensatory filtering. That is, top choices should allow for tradeoffs between key attributes. A further goal is minimizing the number of direct survey questions and emphasizing choice tasks instead.

Conjoint analysis aimed at helping individuals understand themselves differs fundamentally from market-oriented conjoint analysis on behalf of firms in four key respects:



Achieving the objectives and generating the required output given these new forms inputs and analysis poses several challenges.

The Challenge of User Engagement

The entire internet-based exercise must be sufficiently engaging and easy that users do not abandon the website mid-task. The key to keeping users at the site long enough is to make the process fun and engaging with the promise of payoffs such as high-quality recommendations, self-awareness, and a prioritization of key product features and trade-offs.

To make the interview process both efficient and engaging, we employ state-of-art, adaptive questioning methods developed specifically for this new context. A full, 12-attribute, 24-parameter conjoint study can be conducted at the individual level with as few as five to eight vehicle choices (one out of three). This astonishing level of efficiency derives from combining prior population knowledge and a small database of possible utility functions.

In fact the number of database utility functions precisely equals the number of products that can be considered, typically around 3,000 at any given time. Each vehicle has a single utility function associated with it, and that provides enough granularity to match each potential user of the system to a particular favorite vehicle (and other vehicles scoring high using the same utility function. Higher granularity in which each vehicle can have multiple user types, and hence multiple utility functions, is also possible, but will tend to increase the number of choices required to be made by each user during the interview.

Beyond high speed and ease of use, engaging graphics and an appealing theme to the site can make the process "stickier." Questions need to be designed to be very clear with minimal explanation required. Product descriptions and images, and the choices between products need to be intuitive and ultra-clear, traits which can be achieved through extensive pretesting and user-interface analysis. Unlike typical lab-based conjoint studies, this one will mix graphics, animation, text, and allow for some degree of customization for each user. Our development process involved heavy pre-testing of attributes and stimuli before the website was built, then testing of the prototype site, and finally extensive post-testing of the site, clickstream data, and user satisfaction after the site was rolled out.

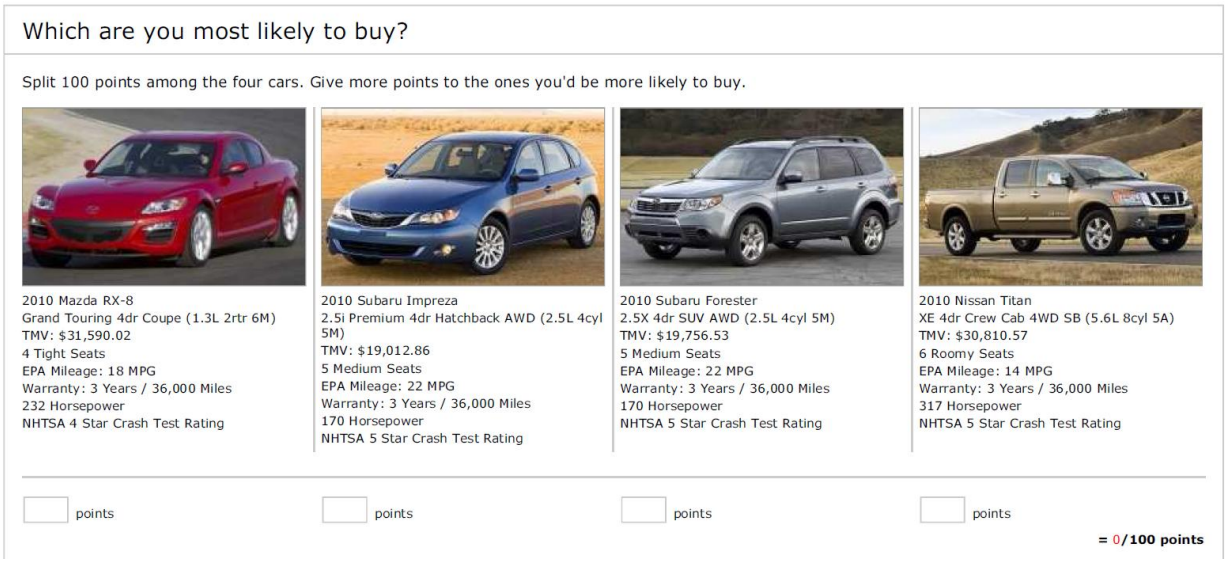
The first attempt at identifying a theme that connects with users was less than successful based on pre-testing. The original theme was to have a cartoon "Professor" help develop a vehicle scoring "formula" on behalf of each user. But this approach backfired as many users interpreted the Professor character as a condescending authority figure who reminded them of being back in school. The responses were reminiscent of the "Clippy" effect, in which Microsoft's helper character became resented.

So a new theme, "AutoMate," was developed in which users would be matched to vehicles similar to the way dating websites match people to each other. This had the added benefit of highlighting the fact that products have utility not only for objective, measurable attributes, but also for subjective, more emotional attributes. In the case of vehicles, these emotional attributes include a vehicle's aesthetics, sporty character, and luxury. To further engage users, the site would provide constant feedback in the form of adaptive attribute priorities and

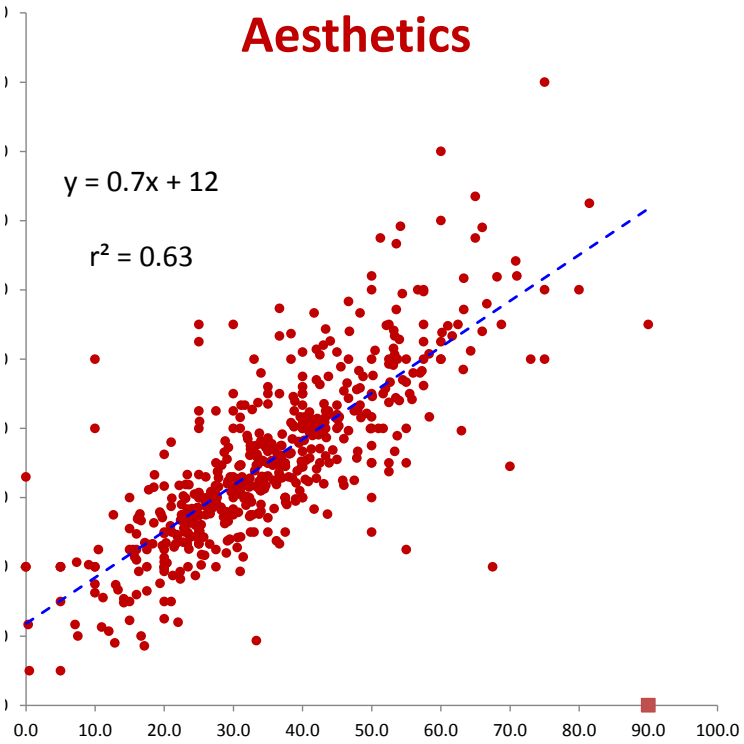
instant gratification in the form of clear vehicle recommendations at the end of the process. The ending payoff features customized product recommendations, detailed facts about each alternative, a personalized report about the user's preferences, and next steps that are easily actionable.

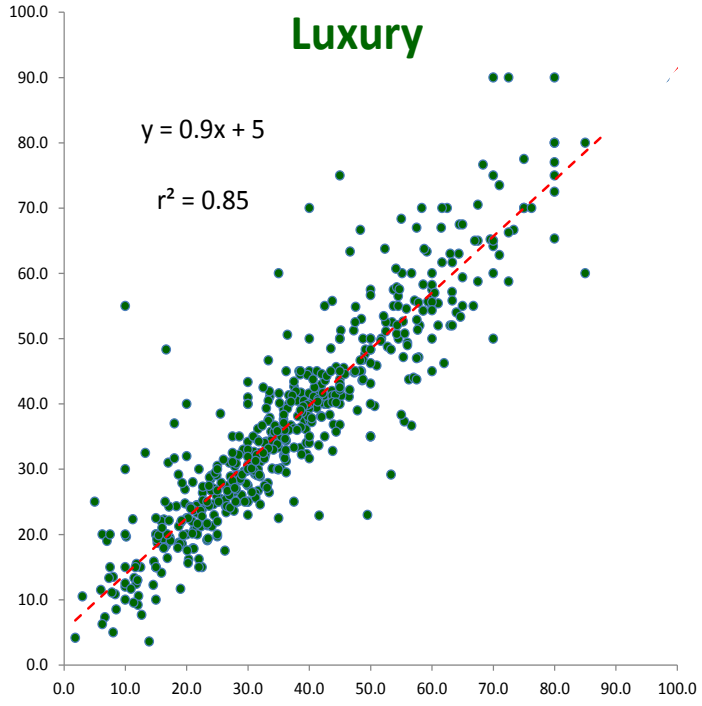
Visual interest is maintained by using attractive background images of the "road less traveled" on the way to self-awareness. And photos of actual vehicles not only make preference measurement more accurate, but are enjoyable to look at during the interview. A progress bar and path history with "Undo" features make navigation more comfortable for users.

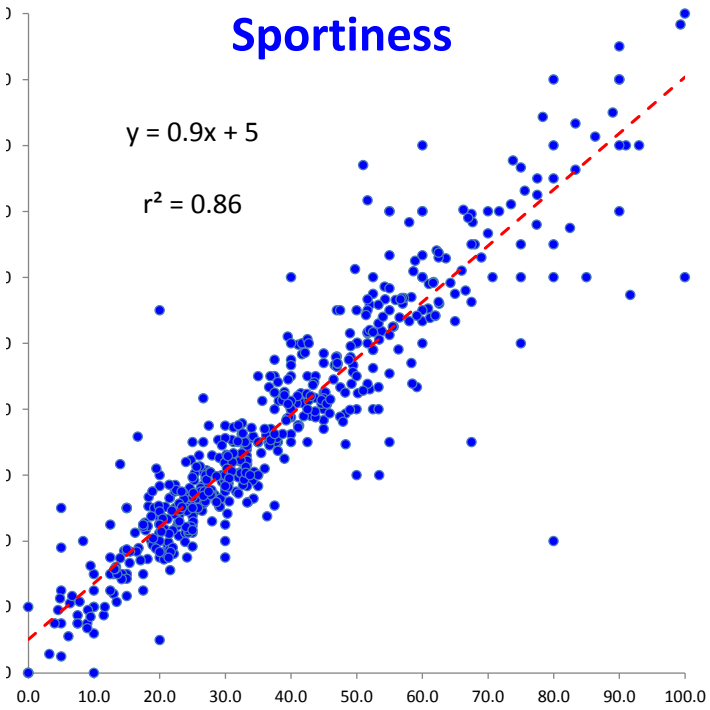
The emotional attributes pose a challenge in that there are no observable, objective measures of people's taste for a product's aesthetics, sportiness and luxury. To solve this dilemma, we rely on the "wisdom of crowds" by conducting a "Preferences of Crowds" contest in which individuals perform constant sum allocations of points to three or four vehicles at a time for each emotional attribute. Then they similarly allocate points, but as best estimates of how other people in the crowd allocated points, on average. The best guessers are rewarded with significant cash prizes.



As the following three graphs demonstrate, the crowd's estimates of emotional attributes are quite accurate, and serve as a good starting point for determining the attribute levels for each of the hundreds of vehicles studied. Each point represents a vehicle's mean allocation of points across all individuals versus their mean estimate of others' point allocations.


Aesthetics
y = 0.7x + 12
r² = 0.63
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0


Luxury
y = 0.9x + 5
r² = 0.85
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0



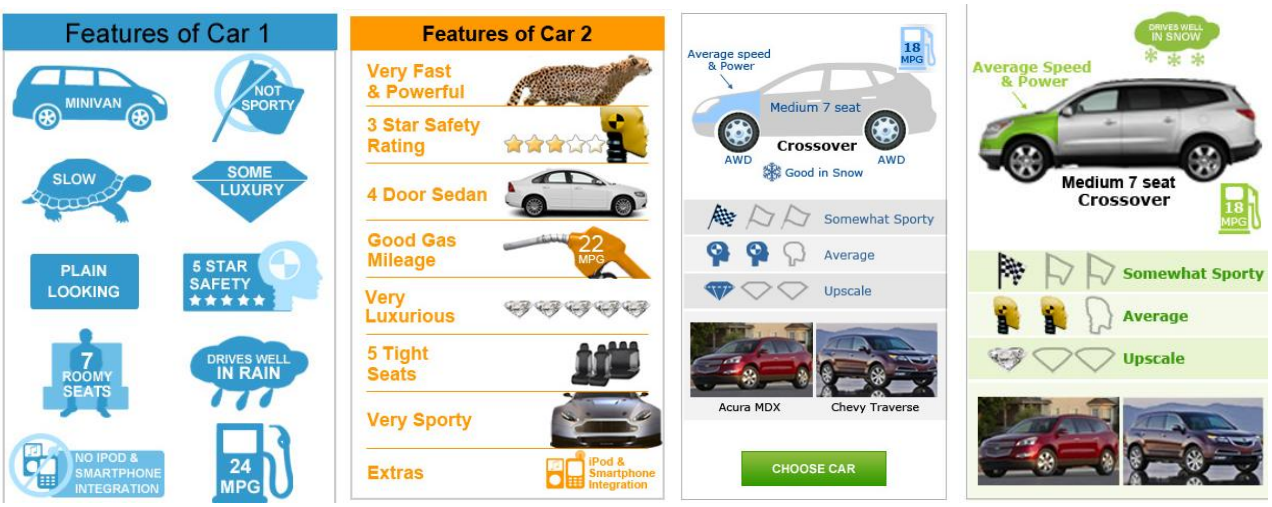
The effort of evaluating thousands of products was split over a large crowd, each member of whom only had to evaluate nine vehicles in total, three-at—time. This experimental design also allows for segmentation based on types of taste.

Vehicles have literally hundreds of possible attributes on which they may be compared. To address the challenge of too many attributes, we combine numerous attributes into "meta-attributes" as depicted on the stimulus shown at right.

We spent a considerable amount of time during pre-testing confirming that potential users would immediately grasp the meaning of each element on the stimuli, and that each would be salient. This is crucial if each attribute is to have its fair chance at being perceived as important to users for whom that attribute actually is important in real buying situations.

In this stimulus, we are also using the image of the vehicle to convey several attributes such as aesthetics and body style, and to reinforce the luxury and sportiness attributes.

The four images below highlight the numerous stimuli formats that were tested with users as a way of determining which were most effective and easiest to understand.



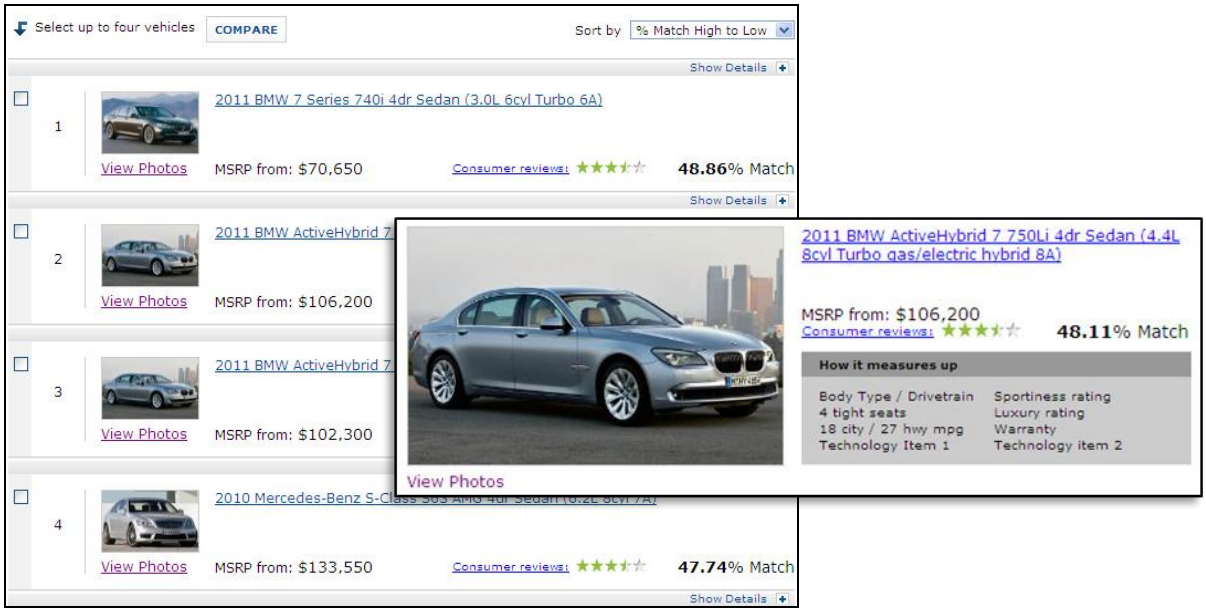
The problem with brands and price

There are 55 brands of vehicles and we can't easily get individual intercepts for each. There are several potential solutions for addressing the brand attribute. For example we could use crowd wisdom as before. Or we could measure brand preference directly using survey questions rather than through choice-based conjoint. Or we could bundle similarly perceived brands to reduce the number of attribute levels. Since none of these solutions is ideal given our other objectives, we choose instead to infer brand preferences from choices by utilizing the population data we have for the entire market. That data allows brand intercepts and covariances to be estimated.

Similarly, the problem with price sensitivity, common to most conjoint studies, is that users aren't spending real money during the interview, so estimated price sensitivity could be inaccurate. Unfortunately, incentive-aligned conjoint analysis is unavailable in the context of vehicle recommendation to millions of potential buyers.

So we propose to leave the price attribute out of the conjoint stimuli and let users self-select their product budget as part of the vehicle recommendation presentation at the end of the process. Vehicle recommendations will be made within various budget constraints.

To that end, we display a pricing grid and let respondents self-select the appropriate budget. We may also be able to infer price sensitivity from observed choices using the population data as we did for brand preferences. A prototype display of product recommendations appears below.



Key Takeaways from This Research:

- Market Conjoint differs from Individual Conjoint, and must therefore be redesigned for this new context.
- Individual measurement is a huge opportunity, but tricky to monetize. Attracting "eyeballs" is one business model. Being rewarded for purchase referrals for which users have opted in may also provide financial reward. Lastly, users may accept a basic report for free and be offered an "upgraded," more detailed version including extra tips for a fee.
- High speed individual measurement poses many challenges, but each has answers. In particular we have addressed the challenges of too many attributes, the need for speed, and the need to make the site "sticky" enough to keep users engaged.
- This may lead to the future of consumer search, as current search methods are extremely efficient, but are quite non-compensatory and not necessarily appropriate for complex product categories involving many tradeoffs.

THE SUCCESS OF CHOICE-BASED CONJOINT DESIGNS AMONG RESPONDENTS MAKING LEXICOGRAPHIC CHOICES

KEITH CHRZAN,
JOHN ZEPP,
AND JOSEPH WHITE
MARITZ RESEARCH

INTRODUCTION

Despite warnings that the format of conjoint tasks can cause response artifacts (Olshavsky and Acito 1980, Wildert 1998), as an industry we have adopted discrete choice experiment (DCE) design technologies that may exacerbate these artifacts, and exacerbate them in a way detrimental to the results of our research.

For example, extremely efficient designs are available for choice experiments (Bunch, Louviere and Anderson 1996, Street and Burgess 2007). One criterion that contributes to an extremely efficient experimental design is "minimal overlap" (Huber and Zwerina 1996). Minimal overlap occurs when the attributes comprising the different alternatives in a choice set have as little duplication in levels as possible. Overlap can be prevented entirely when one has a generic experiment in which all attributes have the same number of levels, and when there are as many alternatives in each choice set as there are levels in each attribute. In such an extremely efficient minimal overlap design, levels for a given attribute will be different for all alternatives in a given choice set.

Statistical efficiency calculations for experimental designs assume that respondents make compensatory choices among alternatives, and that respondents' utility functions are linear (Street and Burgess 2007). Unfortunately, a third or more of respondents may instead be making their choices via non-compensatory lexicographic processes. One study (Campbell *et al* 2006) reports that about 20% of respondents appeared to choose lexicographically. Killi *et al* (2007) report on nine experiments wherein the proportion of respondents appearing to make lexicographic choices ranges from 21% to 41%. Recent models specifically designed to capture non-compensatory choice processes find that up to two-thirds of respondents may be using lexicographic choice rules (Kohli and Jedidi 2007, Yee *et al* 2007).

This combination of highly efficient designs, plus lexicographic choosing, however, may prevent choice responses from being very informative. In many extremely efficient designed experiments, respondents choosing lexicographically supply literally no information about any attributes other than the most important one: the single most important attribute by itself may entirely determine choice. In real markets, multiple products may well be at parity with respect to the most important attributes. In this case, simulations based on utilities derived from respondents' lexicographic choices from choice questions with minimal overlap alternatives may provide no basis for predicting how respondents will choose. Similarly, if respondents choose in a complex manner, say a lexicographic first step followed by compensatory tradeoffs of other attributes, a minimal overlap design may again fail to supply any reliable utilities past those for

the single most important attribute. At stake, therefore is the validity of predictions made from highly efficient experiments conducted on lexicographic choosers.

Liu and Arora (2009) have identified an efficiency calculation for one type of non-compensatory choice model, the conjunctive-disjunctive choice model (Gilbride and Allenby 2006). One could use this efficiency calculation to search for highly efficient designs assuming respondents use conjunctive-disjunctive non-compensatory choice processes. We are not aware of any efficiency calculation under the assumption of lexicographic choosing, so we will compare some alternative design strategies empirically.

In following sections we

a) Define lexicographic choosing

b) Describe how extremely efficient designs minimize or prevent overlap, while other design methodologies do not

c) Illustrate how minimal overlap designs and lexicographic choosing can interact to the detriment of accurate utility estimation

d) Use two empirical studies to test whether lexicographic choosing combines with highly efficient designs to harm utility estimation and prediction, and to investigate whether designs with level overlap built into them ameliorate the harm.

Lexicographic Choosing, Dominant Attributes and Apparently Lexicographic Choosing

Respondents choose lexicographically when they perform an attribute-based culling of alternatives. They make choices sequentially, in order of the importance various attributes have to them. For example, Jones makes choices among alternatives varying on price, brand, feature set and warranty; and he values price most, then brand, then features, then warranty. Jones scans the products arrayed before him and spots the lowest price, whereupon he eliminates all the higher priced products. If but one lowest price product remains, he chooses it, but if two or more products are tied for lowest price, he goes on to consider brand. Among the remaining products he sees one brand he likes more than the others, whereupon he eliminates all the products with the less preferred brands. If only one product remains he chooses it, and if two or more remain under consideration, he moves on to consider feature set. This sequence continues until Jones finds a single product to choose. Note that Jones makes no compensatory tradeoffs at all - in effect, the nth most important attribute is an order of magnitude more important than the (n+1)th most important attribute, and tradeoffs among them do not occur.

Some non-lexicographic choice strategies may appear to be lexicographic in the context of DCEs. DCEs typically feature a relatively small number of alternatives in each choice set and if they have been designed with efficiency in mind, they typically feature minimal overlap. This means that some kinds of behaviors might appear lexicographic, without reflecting true lexicographic preference structures:

- Respondents may have a single dominant attribute for which one level is a “must have,” after which they might well make compensatory choices among attributes. In a design with no overlap, such a dominating attribute would determine choice, but the

respondent's true hybrid lexicographic-compensatory choice process would be invisible in his choice responses.

- Respondents may make compensatory choices, but the levels available to them in the DCE may be so far apart for a given attribute that it becomes dominant. For example, in a study of interior features of an automobile, a DCE might present various features and price points. If the price points are too far apart, however, price may dominate a respondent's decision process and make for apparently lexicographic behavior.
- Respondents may adopt a lexicographic but artificial choice strategy as a shortcut way of navigating the DCE.

Overlap in DCE Design Strategies

Optimally efficient choice experiment designs will tend to have minimal overlap, all else being equal (Huber and Zwerina 1996). Optimal efficiency occurs for designs that minimize the determinant of the Fisher's information matrix, and several strategies produce extremely efficient designs with minimal overlap:

- Shifting strategies based on orthogonal main effects plans (Bunch *et al* 1996)
- Generator strategies based on orthogonal main effects plans (Street and Burgess 2007)
- Designs based on computer search algorithms (Kuhfeld *et al* 1994)

In these designs, if the number of alternatives does not exceed the number of levels for a given attribute, that attribute will not exhibit level overlap in any choice set. Examples of these designs appear as Exhibits 1 to 3, respectively.

Exhibit 1
Highly Efficient Design Without Overlap – Shifting Strategy

Set	Alternative 1				Alternative 2				Alternative 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	3	2	1	2	1	3	2	3	2	1	3	1
2	1	3	2	2	2	1	3	3	3	2	1	1
3	2	2	2	3	3	3	3	1	1	1	1	2
4	3	1	2	1	1	2	3	2	2	3	1	3
5	1	2	3	1	2	3	1	2	3	1	2	3
6	3	3	3	3	1	1	1	1	2	2	2	2
7	2	3	1	1	3	1	2	2	1	2	3	3
8	2	1	3	2	3	2	1	3	1	3	2	1
9	1	1	1	3	2	2	2	1	3	3	3	2

Exhibit 2
Highly Efficient Design Without Overlap – Generator Strategy

Set	Alternative 1				Alternative 2				Alternative 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	1	1	3	3	2	3	2	1	3	2	1	2
2	1	2	1	2	2	1	3	3	3	3	2	1
3	1	3	2	1	2	2	1	2	3	1	3	3
4	2	1	1	1	3	3	3	2	1	2	2	3
5	2	2	2	3	3	1	1	1	1	3	3	2
6	2	3	3	2	3	2	2	3	1	1	1	1
7	3	1	2	2	1	3	1	3	2	2	3	1
8	3	2	3	1	1	1	2	2	2	3	1	3
9	3	3	1	3	1	2	3	1	2	1	2	2

Exhibit 3
Highly Efficient Design Without Overlap – Computer Search Strategy

	Alternative 1				Alternative 2				Alternative 3			
Set	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	2	3	2	3	1	2	3	2	3	1	1	1
2	2	3	1	3	3	2	2	2	1	1	3	1
3	3	1	3	3	2	2	2	1	1	3	1	2
4	2	2	3	2	3	3	1	3	1	1	2	1
5	1	1	2	3	2	3	1	2	3	2	3	1
6	3	3	2	2	2	1	3	3	1	2	1	1
7	3	3	3	1	1	2	2	3	2	1	1	2
8	1	3	3	2	3	2	1	3	2	1	2	1
9	3	1	2	2	1	2	1	3	2	3	3	1

Respondents faced with one of these optimally efficient designs will appear to make lexicographic choices just in case

a) they choose based on true lexicographic preferences

b) they have a hybrid lexicographic-compensatory choice process, but owing to the lack of overlap, only the lexicographic part of their preference structure manifests itself in our DCE

c) their compensatory tradeoffs are invisible because for at least one attribute in our DCE the levels are too far apart and determine choices

d) they choose lexicographically as a way to speed through our questionnaire

Note that in all of these cases, not only do respondents appear to choose lexicographically, but only one attribute, the most important one, completely determines choices. Each respondent

supplies information about a single attribute, and gives us no information whatsoever about any of the other attributes. In the latter three cases other preferences, exist, but they are rendered invisible by the lack of overlap.

Other design strategies do not systematically prevent level overlap. For example, any of the three design strategies above will have overlap for an attribute if we let the number of alternatives in each choice set exceed the number of levels for that attribute. Exhibit 4 shows a design for a 3^4 experiment in nine choice sets of quads.

Exhibit 4
Highly Efficient Quads Design With Overlap

Set	Alternative 1				Alternative 2				Alternative 3				Alternative 4			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	1	1	2	3	2	1	3	1	3	2	3	2	2	3	1	1
2	1	1	1	2	2	2	2	2	3	2	1	1	3	3	3	3
3	1	3	3	1	3	1	2	2	2	2	2	3	3	2	1	1
4	1	2	2	1	2	3	1	2	2	2	1	3	3	1	3	3
5	1	3	2	2	2	1	3	3	3	3	3	1	1	2	1	3
6	2	1	2	1	1	2	3	3	3	1	1	2	1	3	2	3
7	2	3	3	2	1	2	3	1	3	3	1	3	2	1	2	1
8	1	1	1	2	3	3	2	1	2	2	3	2	3	1	2	3
9	2	3	1	3	1	3	3	2	3	2	2	2	1	1	1	1

The Street and Burgess generator design strategy (Street and Burgess 2007) offers the designer the flexibility to control the amount of level overlap present in the design. Careful selection of generators can guarantee that overlap occurs for all attributes or for any subset. Exhibits 5 show a 3^4 triples design generated to overlap on the first attribute only while Exhibit 6 features a 3^4 triples design that has overlap on all attributes. In the empirical studies below, we use this handy feature of the Street and Burgess design strategy to control for overlap in attributes known to dominate choices.

Exhibit 5
Triples Design With Overlap on First Attribute

Set	Alternative 1				Alternative 2				Alternative 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	3	2	1	2	1	3	2	3	3	1	3	1
2	1	3	2	2	2	1	3	3	1	2	1	1
3	2	2	2	3	3	3	3	1	2	1	1	2
4	3	1	2	1	1	2	3	2	3	3	1	3
5	1	2	3	1	2	3	1	2	1	1	2	3
6	3	3	3	3	1	1	1	1	3	2	2	2
7	2	3	1	1	3	1	2	2	2	2	3	3
8	2	1	3	2	3	2	1	3	2	3	2	1
9	1	1	1	3	2	2	2	1	1	3	3	2

Exhibit 6
Triples Overlap on All Attributes

Set	Alternative 1				Alternative 2				Alternative 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	3	2	1	2	1	3	2	3	3	3	1	3
2	1	3	2	2	2	1	3	3	1	1	2	3
3	2	2	2	3	3	3	3	1	2	3	2	1
4	3	1	2	1	1	2	3	2	3	2	2	2
5	1	2	3	1	2	3	1	2	1	3	3	2
6	3	3	3	3	1	1	1	1	3	1	3	1
7	2	3	1	1	3	1	2	2	2	1	1	2
8	2	1	3	2	3	2	1	3	2	2	3	3
9	1	1	1	3	2	2	2	1	1	2	1	1

Some designs devised in the days before the discovery of conditions for optimality also produce level overlap. For example, an L^{MA} design strategy imposes orthogonality both between and within alternatives and will exhibit level overlap (Louviere *et al* 2000). Such a design appears as Exhibit 7 for a 3^4 design in 27 sets of triples.

Exhibit 7
L^{MA} - Orthogonal Between and Within Alternatives

Set	Alternative 1				Alternative 2				Alternative 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	3	3	2	3	3	1	1	2	2	3	2	3
2	2	2	1	2	3	3	1	1	2	3	3	2
3	2	3	3	1	1	1	2	1	1	3	1	3
4	3	1	1	2	1	2	2	2	1	3	3	1
5	3	2	1	1	2	2	2	1	2	2	2	3
6	3	1	3	2	1	3	3	3	2	1	2	3
7	1	2	1	3	1	1	3	1	2	1	1	1
8	1	3	1	2	2	1	3	3	3	3	3	3
9	1	1	2	1	3	3	2	1	3	1	3	3
10	2	3	1	1	1	3	1	3	3	2	2	1
11	3	3	1	3	3	2	2	3	3	1	1	2
12	1	3	3	2	2	2	1	1	1	1	2	2
13	1	3	2	2	2	3	2	2	2	2	1	1
14	1	2	2	3	1	3	2	3	1	3	2	2
15	3	2	3	1	2	3	3	2	3	3	1	2
16	2	2	2	2	3	2	3	3	1	2	1	3
17	2	1	2	3	2	2	3	1	3	3	2	1
18	2	3	2	1	1	2	3	2	2	1	3	2
19	2	1	3	3	2	1	2	3	2	2	3	2
20	1	2	3	3	1	2	1	2	3	2	3	3
21	1	1	1	1	3	1	3	2	1	2	2	2
22	3	2	2	1	2	1	1	3	1	1	3	1
23	3	3	3	3	3	3	3	1	1	2	3	1
24	1	1	3	1	3	2	1	3	2	3	1	1
25	3	1	2	2	1	1	1	1	3	2	1	2
26	2	2	3	2	3	1	2	2	3	1	2	1
27	2	1	1	3	2	3	1	2	1	1	1	3

Adaptive design strategies will also produce level overlap, like Adaptive Choice-Based Conjoint (Johnson 2007) or Tournament-Augmented Conjoint (Chrzan and Yardley 2009). We do not include these design strategies in the current analysis because we want to compare only fixed designs that can be done without specialized software.

Finally, when faced with a large number of attributes, one may opt to use a partial profile design (Chrzan and Elrod 1995). In a partial profile choice set, respondents see only a subset of the attributes with the instruction to assume equivalence in all attributes not shown – in other words, all non-appearing attributes overlap completely. See Exhibit 8 for a 3^7 experiment in 14 partial profile choice sets.

Exhibit 8
Partial Profile Design

	Alternative 1							Alternative2							Alternative 3						
Set	V1	V2	V3	V4	V5	V6	V7	V1	V2	V3	V4	V5	V6	V7	V1	V2	V3	V4	V5	V6	V7
1	0	3	2	0	0	2	0	0	2	3	0	0	3	0	0	1	1	0	0	1	0
2	0	1	1	0	0	2	0	0	3	2	0	0	3	0	0	2	3	0	0	1	0
3	0	1	0	1	0	0	1	0	3	0	2	0	0	3	0	2	0	3	0	0	2
4	0	0	2	0	2	1	0	0	0	3	0	1	2	0	0	0	1	0	3	3	0
5	0	0	0	3	1	0	1	0	0	0	2	2	0	2	0	0	0	1	3	0	3
6	3	0	0	3	0	0	3	1	0	0	2	0	0	1	2	0	0	1	0	0	2
7	1	0	0	0	0	1	3	3	0	0	0	0	2	2	2	0	0	0	0	3	1
8	2	0	1	0	0	0	3	3	0	3	0	0	0	1	1	0	2	0	0	0	2
9	2	0	3	0	0	2	0	3	0	2	0	0	1	0	1	0	1	0	0	3	0
10	0	0	0	3	2	2	0	0	0	0	1	1	3	0	0	0	0	2	3	1	0
11	1	0	3	3	0	0	0	2	0	2	2	0	0	0	3	0	1	1	0	0	0
12	0	0	0	1	0	2	3	0	0	0	3	0	1	1	0	0	0	2	0	3	2
13	0	3	0	0	1	1	0	0	1	0	0	2	3	0	0	2	0	0	3	2	0
14	0	3	0	0	3	0	1	0	1	0	0	1	0	2	0	2	0	0	2	0	3
15	3	0	0	2	1	0	0	1	0	0	1	2	0	0	2	0	0	3	3	0	0
16	0	2	1	2	0	0	0	0	3	3	1	0	0	0	0	1	2	3	0	0	0
17	0	0	0	0	2	2	1	0	0	0	0	3	1	2	0	0	0	0	1	3	3
18	3	3	0	3	0	0	0	2	2	0	1	0	0	0	1	1	0	2	0	0	0
19	2	0	1	0	1	0	0	3	0	3	0	2	0	0	1	0	2	0	3	0	0
20	0	1	3	0	0	0	3	0	2	2	0	0	0	1	0	3	1	0	0	0	2
21	2	3	0	0	2	0	0	3	1	0	0	3	0	0	1	2	0	0	1	0	0

*Zeros in partial profile designs indicate attributes not shown; respondents are instructed that the profiles are all the same on the attributes not shown in a choice set.

Sawtooth Software provides design strategies that allow level overlap, but these are not all so fully explicated that we are able to reproduce them (Sawtooth Software 2008). Note that in any of these design strategies with level overlap, true lexicographic choosers will still choose lexicographically, but that now we should learn more about them. We also stand to learn more about some types of apparently lexicographic choosers.

Does Lexicographic Choosing Harm Utilities Generated from Minimal Overlap Designs?

To demonstrate the potential for trouble when lexicographic choosers meet experimental designs with overlap, we constructed a set of artificial respondents with these utilities that features lexicographic-style dominance on the first attribute:

Table 1 – Simulated Data

Attribute	Level 1	Level 2	Level 3
A	-100	0	100
B	2 (2)	1 (2)	-3 (plug)
C	2 (2)	-1 (2)	-1 (plug)
D	-2 (2)	1 (2)	1 (plug)

Cell=Utility (STD)

The design, a 4^3 shifted triples design with no overlap in three alternatives per choice set, leads to choices wholly determined by attribute A (and specifically by level 3 of attribute A). Analysis via HB-MNL produces the following utilities in Table 2:

Table 2 – HB MNL Data

Attribute	Level 1	Level 2	Level 3
A	-31	-63	94
B	-11	13	-2
C	-14	17	-3
D	8	23	-30

Cell=HB Utility (Actual)

Clearly the analysis picks up valuable information only about level 3 of attribute A. In fact, all other utilities either have reversed signs or they get ordinal relationships wrong (or both). Thus, lexicographic choosing combined with no-overlap discrete choice experimental designs can produce completely useless results.

In the course of running a large number of artificial data sets, however, we found that populations mixing lexicographic and compensatory choosers, or mixed with respect to which attribute or which level dominated choosing often produced estimated utilities that match actual utilities well for lexicographic choosers.

In addition, even homogeneous lexicographic choosers will predict holdouts well if the holdouts contain few overlaps on the dominating attributes. For example, utilities from a homogeneous population of lexicographic choosers will predict holdout triples with no overlap with 100% accuracy (and 50% accuracy when two alternatives tie on the preferred attribute level). This finding suggests that evaluating predictive success depends not only on how well the design strategy captures the respondents' choice strategies, but also on the nature of the holdout simulations. This also suggests both practical and academic caution: practically we want the simulations to be consistent with the design and choice strategies; academically we want to use holdouts that adequately test the design/choice strategy match.

Thus a mismatch between design strategy and respondent choice processes poses a theoretical problem that affects utility estimation and prediction only under certain circumstances. As to whether this potential problem constitutes an actual problem in real world data sets, we turn to two empirical data sets.

EMPIRICAL STUDY 1 – 3^7 LAPTOP COMPUTER STUDY

For the first empirical study, past experience with the market for laptop computers led us to expect possible lexicographic choosing with respect to two attributes, operating system and price (see Exhibit 9 for the attributes and levels used in the study).

Exhibit 9
Attributes and Levels, Laptop Study

Attribute	Level 1	Level 2	Level 3
Operating System (OS)	Windows Vista	Windows 7	Apple/Mac OS X
Processor (CPU)	1.6GHz	2.6 GHz	3.2 GHz
RAM	1 GB	2 GB	4GB
Hard Drive	160 GB	250 GB	500 GB
Screen Size	13"	15"	17"
Battery Life	2 Hours	3.5 Hours	5.5 Hours
Price	600	800	1000

SAMPLE AND ADMINISTRATION

This experiment was part of a commercial study of 1,200 recent purchasers of laptop computers. Respondents each received a random one of four versions of the questionnaire, each using a different experimental design for model estimation but all containing the same six holdout questions. Two holdouts contained overlap on each 0, 2 or 3 profiles. The survey included timers to measure how long respondents took on each of the four versions of the experiment. At the end of the survey, all respondents received three questions used to evaluate their survey experience, to see if the four versions were equally appealing to respondents.

EXPERIMENTAL DESIGN

Again the four versions of the 3^7 experiment employed different strategies with respect to level overlap:

- Version 1: an 18 set shifted triples design with no level overlap
- Version 2: an 18 set generator-based design with level overlap only on the first attribute, operating system, which we suspected from prior research might be subject to lexicographic choosing
- Version 3: an 18 set generator-based triples design with level overlap on the operating system and price attributes
- Version 4: a 21 set partial profile design of triples showing just three attributes per profile; no overlap occurred among visible attributes, but the questions instructed the respondents that profiles were identical on all attributes not shown (i.e. complete overlap across the three profiles).

MODELING

We used the Sawtooth Software CBC/HB program for hierarchical Bayesian MNL (Sawtooth Software 2008) to estimate simple main effects models for all four versions of the experiment. Results using an aggregate MNL model were similar enough to require no separate discussion.

HOLDOUTS

The design for the six holdout questions appears in Exhibit 10.

Exhibit 10
Holdout Design, Laptop Study

Set	Alternative A							Alternative B							Alternative C						
	V1	V2	V3	V4	V5	V6	V7	V1	V2	V3	V4	V5	V6	V7	V1	V2	V3	V4	V5	V6	V7
1	1	2	1	0	0	1	0	2	0	2	1	1	2	1	0	1	0	2	2	0	2
2	0	2	1	2	1	2	1	2	1	0	1	0	1	0	1	0	2	0	2	0	2
3	2	2	2	2	0	0	1	0	0	0	0	1	1	2	0	1	1	1	2	2	0
4	1	1	1	1	2	2	2	1	2	2	2	0	0	1	0	0	0	0	1	1	2
5	1	2	1	2	1	2	1	1	0	2	0	2	0	2	1	1	0	1	0	1	0
6	2	2	1	2	1	2	0	2	1	0	1	0	1	0	2	0	2	0	2	0	0

RESULTS

We assess the quality of results using two imperfect measures of predictive validity – in-sample holdout hit rates (the ability of respondents' own utilities to predict their responses to holdout questions they themselves answered) and out of sample mean absolute errors (the ability of respondents' utilities to predict choice shares of holdout choices given to other respondents).

Table 3 – In-Sample Holdout Hit Rates

		Design strategy		
Holdout Type	Efficient	Overlap OS	Overlap OS & Price	Partial Profile
No Overlap	80	81	**74***	**70***
Partial overlap on OS only	81	83	78	78
Partial overlap on OS & Price	67	73	71	62
Complete overlap on OS	67	**74***	67	72
Complete overlap on OS & Price	65	71	**73***	**80***
Overall	73	**77**	73	72

*significantly different from efficient design at p<.05.

Designs that pay too little attention to the ill effects of dominance (the efficient non-overlap design) suffer when holdouts contain overlaps. On the other hand, designs that contain too much overlap (on both operating system and price or partial profile designs) suffer when holdouts contain no overlap.

Across a broad range of holdouts, the design with overlap on operating system, the attribute must suspected to dominate choices in a lexicographic manner, garners the highest hit rates. In the extreme case of complete overlap on operating system and price among the alternatives in holdout choices, however, the partial profile design performs best (only 15 of the 21 choice sets included either or both operating system and price so even if they dominated choices, other choice sets enabled the design to capture preferences for other attributes).

Table 4 – Out-of-Sample Holdout MAE

	Design strategy			
Holdout Type	Efficient	Overlap OS	Overlap OS & Price	Partial Profile
No Overlap	3.0	4.8	5.9	7.1
Partial overlap on OS only	4.8	0.6	3.4	4.0
Partial overlap on OS & Price	7.5	2.9	5.3	6.3
Complete overlap on OS	12.8	3.7	7.4	7.0
Complete overlap on OS & Price	12.2	5.0	7.3	4.3
Overall	7.2	3.6	5.9	4.3

Prediction suffers when mismatches between design and holdout levels of overlap diverge, with the degradation in the ability of the efficient non-overlap design's ability predict holdouts with high levels of overlap standing out especially. The design targeting overlap only on operating system performs best across a range of holdouts, but the non-overlap efficient design does the best job of predicting non-overlap holdouts and the partial profile model does the best job of predicting holdouts with the most extensive level of overlap.

EMPIRICAL STUDY 2 – 3^4 BOXED LUNCH STUDY

The second empirical study, focused on choices of boxed lunches, benefitted from 20+ pretests which showed evidence of lexicographic choosing with respect to two of the attributes, sandwich type and side item (see Exhibit 11 for a full list of attributes and levels).

Exhibit 11
Attributes and Levels for Boxed Lunch Study

Attribute	Level 1	Level 2	Level 3
Sandwich	Turkey	Roast beef	Veggie wrap
Side	Chips	Fruit	Cole slaw
Dessert	Brownie	Jello	Cookie
Drink	Tea	Soda	Water

SAMPLE AND ADMINISTRATION

A convenience sample of 396 students at two Indiana universities completed a brief three page paper and pencil survey (we discarded another 45 incomplete surveys). Respondents each received randomly one of four versions of the questionnaire. Each version contained a different experimental design for model estimation but all versions contained the same six holdout questions. Two holdouts contained overlap on each 0, 2 or 3 profiles.

EXPERIMENTAL DESIGN

The four versions of the 3^4 experiment employed different strategies with respect to level overlap:

- Version 1: a shifted triples design with no level overlap
- Version 2: a generator-based triples design with level overlap only on the first attribute (sandwich type), which was subject to lexicographic choosing based on the results from the pretests
- Version 3: a generator-based triples design with level overlap all attributes
- Version 4: a generator based quads design with minimal (but necessarily some) overlap on all attributes

MODELING

We ran both aggregate MNL and HB-MNL models of simple main effects models for all four versions of the experiment and the results were again similar enough that we need only show the HB results. Respondents identified best and worst alternatives in each choice set, supporting estimation of a best-worst discrete choice experiment (Louviere *et al* 2008).

HOLDOUTS

The design for the six holdout questions appears in Exhibit 12.

Exhibit 12
Holdout Design for Boxed Lunch Experiment

Set	Alternative 1				Alternative 2				Alternative 3			
	V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
1	1	2	1	2	1	3	2	3	1	1	3	1
2	3	1	1	2	1	2	2	3	2	3	3	1
3	2	1	2	3	3	2	3	1	1	3	1	2
4	2	1	2	1	3	2	3	2	3	3	1	3
5	1	2	3	1	1	3	1	2	3	1	2	3
6	2	3	3	3	2	1	1	1	1	2	2	2

RESULTS

Table 5 shows within sample prediction – the ability of HB utilities generated from first-choice models of the four design types to predict holdout choices with different amounts of overlap.

Table 5 – In-Sample Holdout Hit Rates

	Design strategy			
Holdout Type	Efficient triples	Triples – overlap on sandwich	Triples – overlap on all attributes	Minimum overlap quads
No Overlap	75	77	70	74
Partial overlap on sandwich	74	75	71	69
Complete overlap on sandwich	62	68	66	66
Overall	72	74	70	70

As above, the design targeting overlap only on the attribute suspected in advance to dominate, sandwich type, performs well across the range of holdouts and again, mismatches between design and holdout levels of overlap result in lower prediction accuracy.

Table 6 shows the ability of the four models to predict holdout choices among out-of-sample respondents (among the respondents who did not receive the modeled design). Results appear as mean absolute errors or prediction

Table 6 – Out-of-Sample MAE

	Design strategy			
Holdout Type	Efficient triples	Triples – overlap on sandwich	Triples – overlap on all attributes	Minimum overlap quads
No Overlap	3.3	4.4	6.3	5.2
Partial overlap on sandwich	5.1	2.2	3.5	3.2
Complete overlap on sandwich	5.7	4.0	5.3	4.1
Overall	4.6	3.3	4.7	4.0

As in the laptop study reported above, out-of-sample predictions fare best when the design targets overlap on the dominant attribute, unless overlap does not appear in holdouts.

Table 7 – Out-of-Sample MAE (first choice/BW-DCE)

	Design strategy			
Holdout Type	Efficient triples	Triples – overlap on sandwich	Triples – overlap on all attributes	Minimum overlap quads
No Overlap	3.3/4.5	4.4/3.2	6.3/6.5	5.2/5.1
Partial overlap on sandwich	5.1/4.9	2.2/2.5	3.5/2.6	3.2/2.6
Complete overlap on sandwich	5.7/3.1	4.0/1.9	5.3/0.8	4.1/3.6
Overall	4.6	3.3	4.7	4.0

Plausibly, the best-worst questions perform better because there are two of them per choice set. To counteract this we built another best-worst model that used only half of each respondent's questions, thus equalizing respondent effort, in terms of the number of keystrokes required for the task. When we did so, the best-worst questions still outperform first choice questions in predicting holdouts with extensive overlap. Best-worst appears to get its strength from the nature of the task itself and not just from requiring respondents to provide more responses.

DISCUSSION

In the course of this research, our appreciation for the effects of lexicographic choosing on design overlap and holdout selection deepened. One can easily show with simulations that HB or MNL analysis will do a poor job of reproducing respondent utilities if respondents make lexicographic or dominated choices. Unfortunately, very many, perhaps most, respondents appear to make lexicographic choices, or at least they allow a single attribute to dominate their choices. When this happens, experimental designs that contain overlaps, ideally on the attributes that dominate choice, will produce more informative utilities for the non-dominating attributes, resulting in greater ability to predict difficult overlap-containing holdout choices and presumably in actual markets containing highly competitive products.

That our results depend to some extent on the details of how much overlap occurs in holdouts (which is why we used a range of overlap amounts in our holdouts) suggests that, as an industry, we have paid inadequate attention to the construction of holdout choices. A welcome addition to the literature on choice model validation using holdout choices would be a theory about how holdouts should be constructed. At a minimum, analysts reporting success in predicting holdout choices should identify the attribute levels defining their holdout alternatives and choice sets.

Finally, our research supports, to some extent, the collection of both best and worst choices for choice sets containing three or more alternatives. Models built from best and worst choices further improved prediction when holdouts contained extreme levels of overlap (i.e. when they represented highly competitive markets).

REFERENCES

Bunch, David S., Jordan J. Louviere and Don Anderson (1996) "A Comparison of Experimental Design Strategies for Multinomial Logit Models: The Case of Generic Attributes." Technical Report, University of California, Davis.

Campbell, Danny, W. George Hutchinson and Riccardo Scarpa (2006) "Lexicographic Preferences in Discrete Choice Experiments: Consequences on Individual-Specific Willingness to Pay Estimates," *Nota Di Lavordo*, **128**.

Chrzan, Keith and Terry Elrod (1995) "Partial Profile Choice Experiments: A Choice-Based Approach for Handling Large Numbers of Attributes," paper presented at the AMA's Advanced Research Techniques Forum, Monterey, CA.

Chrzan, Keith and Daniel Yardley (2009) "Tournament-Augmented CBC," *Sawtooth Software Conference Proceedings*, Sequim: Sawtooth Software, in press.

Johnson, Rich and Bryan Orme (2007) "A New Approach to Adaptive CBC," *Sawtooth Software Conference Proceedings*, Sequim: Sawtooth Software, 85-109.

Gilbride, Timothy and Greg M. Allenby (2006) "Estimating Heterogeneous EBA and Economic Screening Rule Choice Models," *Marketing Science*, **25**, 494-509.

Green, Paul E. and Yoram Wind (1973). *Multiattribute Decisions in Marketing: A Measurement Approach*. Hinsdale: Dryden.

Huber, Joel, Klaus Zwerina (1996) "The Importance of Utility Balance in Efficient Choice Designs," *Journal of Marketing Research*, **33**, 307-317.

Killi, Marit, Ase Nossum and Knut Veisten (2007) "Lexicographic Answering in Travel Choice: Insufficient Scale Extensions and Steep Indifference Curves?" *European Journal of Transport and Infrastructure Research*, **7**, 39-62.

Kohli, Rajeev and Kamel Jedidi (2007) "Representation and Inference of Lexicographic Preference Models and Their Variants," *Marketing Science*, **26**, 380-399.

Kuhfeld, W.F., Tobias, R.D., and Garratt, M. (1994), "Efficient Experimental Design with Marketing Research Applications," *Journal of Marketing Research*, **31**, 545–557.

Liu, Qing and Neeraj Arora (2009) "Efficient Choice Designs Under Non-Compensatory Models," paper presented at the INFORMS Marketing Science Conference, Ann Arbor.

Louviere, Jordan J., David A. Hensher and Joffre D. Swait (2000) *Stated Choice Methods: Analysis and Application*. Cambridge: Cambridge University.

Louviere, Jordan J., Deborah Street, Leonie Burgess, Nada Wasi, Towhidul Islam and Anthony A. J. Marley (2008) "Modeling the Choices of Individual Decision-Makers by Combining Efficient Choice Experiment Designs with Extra Preference Information," *Journal of Choice Modeling*, **1**, 126-163.

Lusk, Jayson L. and F. Bailey Norwood (2005) "Effect of Experimental Design on Choice-Based Conjoint Valuation Estimates," *American Journal of Agricultural Economics*, **87**, 771-785.

Olshavsky, Richard W. and Franklin Acito (1980) "The Impact of Data Collection Procedure on Choice Rule," in *Advances in Consumer Research Volume 7*, ed. Jerry C. Olson. Ann Arbor: Association for Consumer Research, 729-732.

Orme, Bryan (2009) "Fine-Tuning CBC and Adaptive CBC Questionnaires," retrieved from http://sawtoothsoftware.com/techpap/finetune.pdf.

Sandor, Zsolt and Michel Wedel (2002) "Profile Construction in Experimental Choice Designs for Mixed Logit Models," *Marketing Science*, **21**, 455-475.

Sandor, Zsolt and Michel Wedel (2005) "Heterogeneous Conjoint Choice Designs," *Journal of Marketing Research*, **42**, 210-218.

Sawtooth Software (2008). *The CBC System for Choice-Based Conjoint Analysis*, Sequim: Sawtooth Software.

Street, Deborah J. and Leonie Bugess (2007) *The Construction of Optimal Stated Choice Experiments: Theory and Methods*, Hoboken: Wiley.

Wildert, Staffan (1998) "Stated Preference Studies: The Design Affects the Results," in *Travel Behavior Research: Updating the State of Play*, eds. Juan de Dios Ortuzar, David Hensher and Sergio Hara-Diaz, 105-121.

Yee, Michael, Ely Dahan, John R. Hauser and James Orlin (2007) "Greedoid-Based Noncompensatory Inference," *Marketing Science*, **26**, 532-549.

MENU-BASED CHOICE MODELING USING TRADITIONAL TOOLS

BRYAN ORME
SAWTOOTH SOFTWARE, INC.

EXECUTIVE SUMMARY

Menu-Based Choice (MBC) research studies are becoming more popular as businesses implement *mass customization* sales models. The good news for conjoint researchers is that the existing tools for designing and analyzing CBC studies may also be used for MBC. The not-so-good news is that until software is developed to automate many of the steps (and we're working on that), the process is typically more challenging and time-consuming than for CBC studies.

Both traditional "fixed" and randomized design strategies are quite effective for designing MBC questionnaires. If randomized designs are used, the intuitive counting analysis approach can convey top-line results, including detailed price sensitivity curves. Multinomial logit (MNL) may be used in analysis, including aggregate logit, latent class, and hierarchical Bayes (HB) variants. Simulators may be built with Excel.

BACKGROUND

Increasingly, stated preference choice projects involve Menu-Based Choice scenarios (MBC) where respondents can select from one to multiple options from a menu. This is not surprising, given the fact that buyers are commonly allowed to customize products and services (mass customization). Examples include choosing options to put on an automobile, designing employee benefits packages, combination drug therapy choices for pharma, selections from a restaurant menu, banking options, configuring an insurance policy, or purchasing bundled vs. *a la carte* services including mobile phones, internet, and cable.

Here is a very simple menu, where the respondent chooses options and a total price is shown:

Which of the following options would you buy? Select as many as you wish, or none of the items.

☑ Option A $12
❑ Option B $24
☑ Option C $7
❑ Option D $55
☑ Option E $3

Total Price of Selected Options: __$22__

Figure 1

Respondents can select between zero to five choices on the menu in Figure 1. This particular respondent has selected three items, for a total price of $22. There are 2^5=32 possible ways respondents can complete this menu. The prices shown might always stay the same, or perhaps the questionnaire is designed so that some or all of the prices vary between respondents, or even across repeated menus given to the same respondent.

If prices are varied across menu questions, we can observe whether changing the prices influences what respondents pick. Economic theory suggests that as the price of a menu item *increases* its likelihood of choice will *decrease*. How strong is that relationship? Is it fairly linear? Does reducing the price for an item cause a different item on the menu to be more likely (or even *less* likely) to be chosen? In other words, are items on the menus *substitutes* or *complements*? Menu-Based Choice (MBC) experiments can investigate such issues.

MBC questionnaires are often used to investigate bundling vs. *a la carte* strategies. We've been showing a case study involving fast-food menu choices at Sawtooth Software's advanced CBC trainings for over five years now:

<table>
<tr><td colspan="3">Menu Scenario #1: Please imagine you pulled into a fast-food restaurant to order dinner for <u>just yourself</u>. If this were the menu, what (if anything) would you purchase?</td></tr>
<tr><td>❑ Deluxe Hamburger Value Meal
-Deluxe Hamburger
-Medium fries
-Medium drink

$3.99</td><td>❑ Chicken Sandwich Value Meal
-Chicken Sandwich
-Medium fries
-Medium drink

$5.59</td><td>❑ Fish Sandwich Value Meal
-Fish Sandwich
-Medium fries
-Medium drink

$3.99</td></tr>
<tr><td colspan="2">(Only order sandwiches, fries or drinks from this area if you did not pick a value meal above.)

Sandwiches:
❑ Deluxe Hamburger $1.99
❑ Chicken Sandwich $3.59
❑ Fish Sandwich $1.99

Fries:
❑ Small $0.79
❑ Medium $1.49
❑ Large $1.69

Drinks:
❑ Small $0.99
❑ Medium $1.69
❑ Large $2.19</td><td>Salads:
❑ Cobb dinner salad $4.79
❑ Grilled chicken salad $4.39

Healthy Sides:
❑ Carrots/Celery with Ranch dressing $1.19
❑ Apple slices/Grapes with dipping sauce $0.99

Desserts:
❑ Apple/Cherry/Berry pie $0.99
❑ Cookies $1.19</td></tr>
<tr><td colspan="3">❑ I wouldn't buy anything from this menu.
I'd drive to a different restaurant, or do something else for dinner.</td></tr>
</table>

Figure 2

Past Literature and Sawtooth Software Presentations

A number of articles have been published on menu-based choice, including:

- Ben-Akiva, M. and S. Gershenfeld (1998), "Multi-featured Products and Services: Analysing Pricing and Bundling Strategies," Journal of Forecasting, 17.
- Liechty, J., Ramaswamy, V., and S. Cohen (2001), "Choice-Menus for Mass Customization: An Experimental Approach for Analyzing Customer Demand with an Application to a Web-based Information Service," JMR, 39 (2).
- Cohen, S. and J. Liechty, (2007), "Have it Your Way: Menu-based conjoint analysis helps marketers understand mass customization," Marketing Research, 19:3.

The following was presented at the ART/Forum (American Marketing Association):

- Conklin. M., B. Paris, T. Boehnlien-Kearby, C. Johnson, K. Juhl, A. Zanetti-Polzi, K. Gustafson, B. Palmer, (2007) "Menu Based Choice Models," ART Forum.

And, the Sawtooth Software Conference has also seen some useful papers on MBC:

- Bakken, David and Len Bayer (2001), "Increasing the Value of Choice-Based Conjoint with 'Build Your Own' Configuration Questions," Sawtooth Software Conference Proceedings, pp 99-110.
- Bakken, David and Megan Kaiser Bond (2004), "Estimating Preferences for Product Bundles vs. *a la carte* Choices," Sawtooth Software Conference Proceedings, pp 123-134.
- Johnson, Richard, Bryan Orme and Jon Pinnell (2006), "Simulating Market Preference with 'Build Your Own' Data," Sawtooth Software Conference Proceedings, pp 239-253.
- Rice, Jennifer and David Bakken (2006), "Estimating Attribute Level Utilities from 'Design Your Own Product' Data—Chapter 3," Sawtooth Software Conference Proceedings, pp 229-238.

David Bakken's papers have been especially useful, and some of the ideas we present here are drawn directly from his work.

Designing MBC Studies

For traditional conjoint and CBC, the focus has been on designing a set of product concepts that respondents rate or choose. With MBC studies, the focus is on asking respondents to *configure their preferred choice* by making from zero to multiple selections from a menu of possible selections. Respondents are permitted to take a more *proactive* approach in designing appropriate products in MBC, whereas they tend to be placed in a more *reactive* stance with CBC questionnaires.

In conjoint analysis, we consider multiple factors (attributes), where each attribute has at least two levels. Menu-based choice problems also involve multiple factors, each having multiple levels. Whereas we often think of a CBC question as being composed of multiple product concepts (cards), we should think of the entire MBC menu question being controlled by a *single*

card. This allows researchers to use the familiar tools for conjoint design with MBC experiments (e.g. CBC or CVA software, Warren Kuhfeld's SAS routines), except that the number of factors for MBC experiments will often be larger than for traditional conjoint or CBC.

In the MBC studies we've designed at Sawtooth Software, we've done web-based surveys and have used CBC's Complete Enumeration, Balanced Overlap, or Shortcut design strategies. These are randomized design routines that explicitly control for level balance, and in the case of Complete Enumeration and Balanced overlap, also control for orthogonality. Each respondent is randomly selected to receive one of many versions of the design, where each version has been carefully constructed. However, even purely random design strategies, such as would be available using randomized list functions available in many web interviewing systems (including SSI Web), will produce robust designs that work quite well for MBC studies.

Sample Study

Early in 2010, we conducted a methodological research study among approximately 1600 respondents (pre-screened for intention to purchase a new car in the next few years). 800 respondents were used for building the MBC models (calibration respondents), and 800 respondents were used as holdout sample. We used Western Wats' panel for respondent recruitment and invitations, and fielded the study using our SSI Web platform. (We thank our colleagues at Western Wats for their support of our R&D efforts and for their excellent service.)

Our MBC study actually consisted of two separate MBC exercises (shown below), which we analyzed independently.

At the beginning of the survey, we asked respondents how much they expected to pay for their next new vehicle, and to rate their preferences for automobile options (to be used as covariates in HB modeling). Next, we asked respondents to select the three new vehicles they were most likely to consider purchasing, and to indicate for each how much they expected to pay. The vehicle choices were provided in drop-down lists, developed using information from: http://www.automotive.com/new-cars/index.html. All respondents provided three vehicle choices in their consideration set.

In MBC Exercise 1, we showed the vehicle that respondents picked as their top considered vehicle, at a base price $2,000 *less than* the amount they said they were expecting to pay. The exercise consisted of eight choice tasks like the one directly below:

Let's assume you were going to purchase the **Honda Accord** and it didn't have any of the options below as standard features. If the prices for the options were as shown below, which options would you add to your vehicle?

(If you would add no options, just click the "Next" button)

Base Price: $23,000

☐	$1,500	Alloy Wheels
☐	$900	Moonroof/Sunroof
☐	$300	XM Radio (+ $13/month)
☐	$800	Leather Seats
☐	$350	Security System
☐	$1,000	Backup/parking assist sensor with rearview camera
☐	$600	Hands-Free Phone System
☐	$1,300	Navigation system (in dash)

Total: $23,000

Figure 3

The prices for the different options were varied (within respondent) using an experimental design (CBC's Complete Enumeration design method). Four price levels were varied per option, as shown below:

	Price 1	Price 2	Price 3	Price 4
Alloy Wheels	$1,500	$1,750	$2,000	$2,500
Sunroof	$500	$700	$900	$1,200
XM Radio	$300	$400	$500	$600
Leather Seats	$600	$800	$1,000	$1,200
Security System	$150	$200	$250	$350
Parking Assist	$600	$700	$800	$1,000
Hands-Free Phone	$400	$500	$600	$800
Navigation System	$1,000	$1,300	$1,600	$2,000

Figure 4

Exercise 2 was a bit more complex. We showed *all three* top considered vehicles. Respondents were asked to choose one of the vehicles, and then to add any options to that chosen vehicle. Respondents completed eight tasks as shown in Figure 5:

If you were deciding between the following three cars, and the prices were as shown, which car would you select, and which options would you add to it?

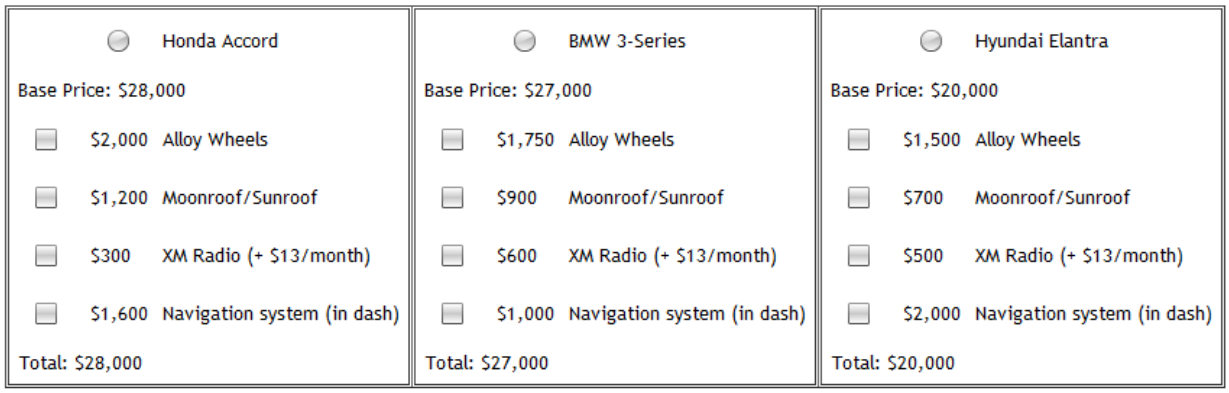


Figure 5

The four options (Alloy Wheels, Moonroof, XM Radio, and Navigation System) were each varied over four prices, as was shown in Figure 4. This time, we used CBC's Balanced Overlap design strategy to generate the experimental design (as some level overlap in the design would support estimation of cross-effects better than Complete Enumeration's minimal overlap strategy). Additionally, the base price of the vehicle was varied over three price points: $3,000 less than expected price, expected price, and $3,000 more than expected price.

Counting Analysis

Randomized designs are not only a robust and straightforward way to design complex MBC tasks, but they permit a simple form of top-line analysis: counts. With counting analysis, we simply compute the percent of times that an option was chosen. We can count that likelihood of choice overall, or split out by various prices shown on the menu.

Exercise 1

For Exercise 1, where respondents configured their top-most considered vehicle, we've summarized the percent of times options were chosen (across all price variations) in the table below. The most commonly selected item was security system (selected for 67% of the configured vehicles).

Exercise 1
Counting Results: Choice of Menu Options

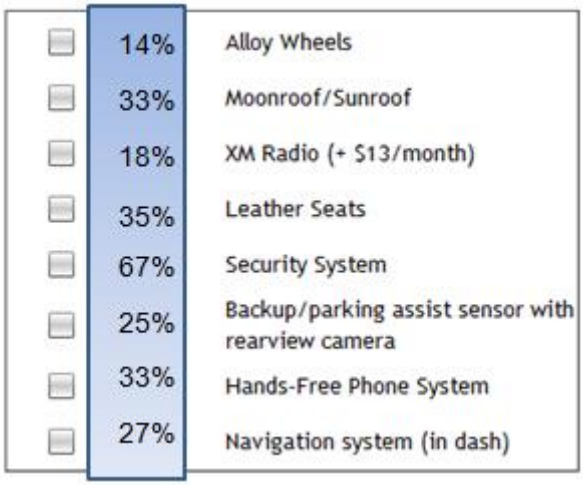


Figure 6

We can also count the percent of times each item was chosen when offered at each of its four price points, and plot the results in a chart. We've also performed a simple log-log regression to compute the price elasticity of demand for each price curve (shown in parentheses after each item's label).

Counting Results: Options by Price

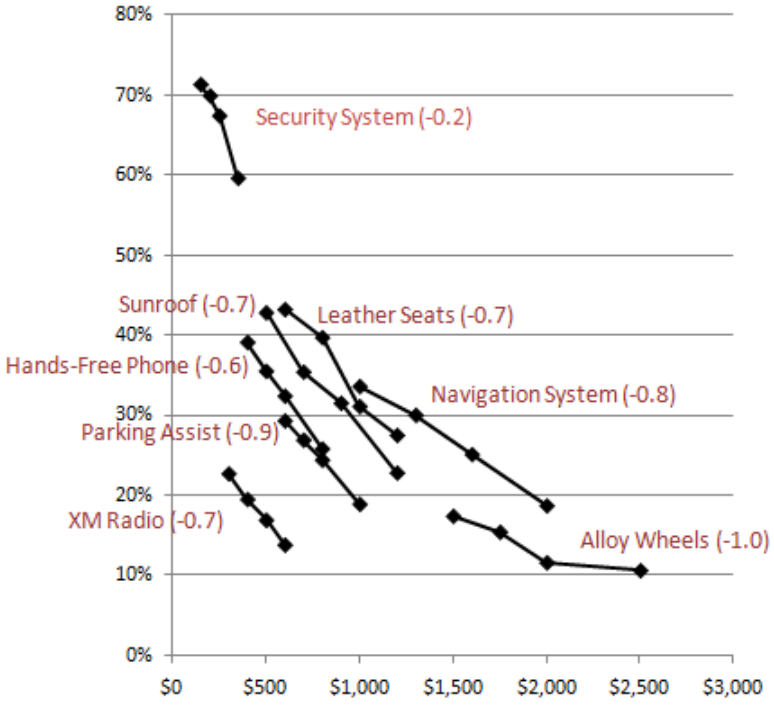


Figure 7

Each of the charted proportions in Figure 7 has a base size of 800 respondents x 8 tasks / 4 price points = 1,600. In other words, each item was available to be chosen at each price point 1,600 times. Because each price point occurred about an equal number of times with every other item's price in the design, we can summarize the independent price effect for each item using counting analysis. Assuming the worst-case scenario 50% proportion, the margin of error (95% confidence) is +/- 2.4% for each estimate. Figure 7 contains a lot of information, and the data are quite precise (+/- 2.4%), given the 1,600 data points supporting each proportion. As you can see, simple counting analysis can relay quite intuitive and useful information.

We may also count the *combinations* of items that were configured in the menu. With 800 respondents x 8 choice tasks, there were 6,400 total vehicles that were configured. We can tally the combinations selected, and report the top 10 most configured options (Figure 8).

Counting Results: 10 Most Common Combinations Selected

Alloy Wheels	Sunroof	XM Radio	Leather Seats	Security System	Parking Assist	Hands-Free Phone	Navi-gation System	%
				✓				12.0%
								8.9%
	✓			✓				3.8%
			✓	✓				3.5%
				✓		✓		3.1%
	✓		✓	✓				2.3%
				✓	✓			2.2%
			✓					2.2%
	✓							2.0%
	✓			✓		✓		1.8%
							Total:	41.7%

Figure 8

The most common configuration across the 6,400 tasks was to add only Security System (12.0% of the menus completed resulted in this choice). If two options were chosen, the most common selection was Sunroof and Security system, with 3.8% of the choices. The top 10 most common configurations account for 41.7% of the total choices made.

We also employed counting analysis to examine the cross-effects among menu items. For example, what was the effect of the price of the Security System on the choice for Sunroof? It turns out that *no* cross effects were significant! This surprised us, and makes Exercise 1 a very easy data set (but somewhat boring) to model.

Exercise 2

Exercise 2 is more complicated, since there were three vehicles shown, with the base price of the vehicle as well as option prices varying (within respondent). The base prices of the top three considered vehicles were varied by -$3,000 to +$3,000 of expected price. The summary counts for items selected on the menu (across all price manipulations) are:

Exercise 2
Counting Results: Choice of Menu Options

If you were deciding between the following three cars, and the prices were as shown, which car would you select, and which options would you add to it?

47.2% 1st Considered Car		29.5% 2nd Considered Car		23.4% 3rd Considered Car	
Base Price: $28,000		Base Price: $27,000		Base Price: $20,000	
9.2%	Alloy Wheels	5.2%	Alloy Wheels	4.6%	Alloy Wheels
17.8%	Moonroof/Sunroof	11.2%	Moonroof/Sunroof	8.5%	Moonroof/Sunroof
9.3%	XM Radio (+ $13/month)	6.2%	XM Radio (+ $13/month)	4.7%	XM Radio (+ $13/month)
15.4%	Navigation system (in dash)	9.4%	Navigation system (in dash)	7.6%	Navigation system (in dash)

Figure 9

The price changes on the base price of the vehicles (-$3,000 to +$3,000) had a very large effect on the choice likelihood for the vehicle (Figure 10). If the price of the top considered vehicle was increased by $3,000, respondents were only about 20% likely to select that vehicle, and nearly 80% likely to switch to one of the other two vehicles.

Counting Analysis: Likelihood Choosing Vehicle Due to Changes in Base Price

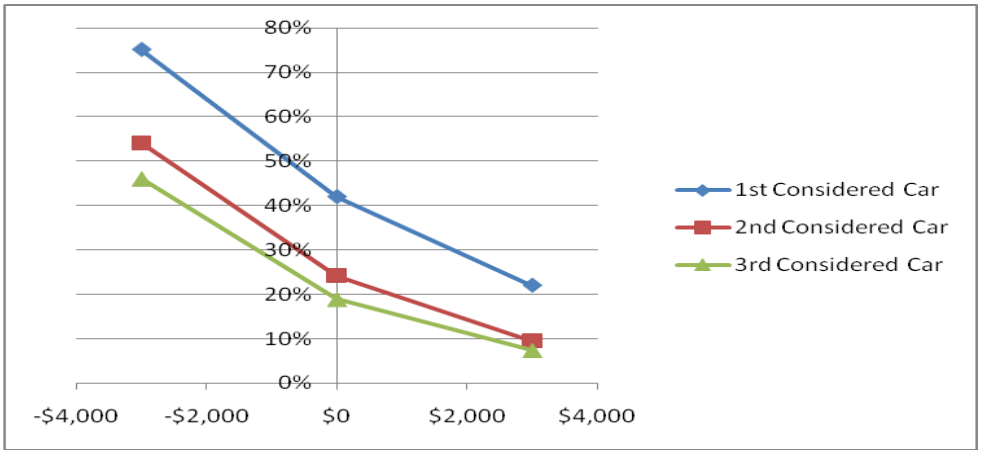


Figure 10

Given that respondents picked a particular vehicle, we can count the likelihood of selecting the four options (Alloy Wheels, Sunroof, XM Radio, Navigation System) at each of their prices.

Counting Analysis: Choice of Option x Prices, Given the Choice of Vehicle

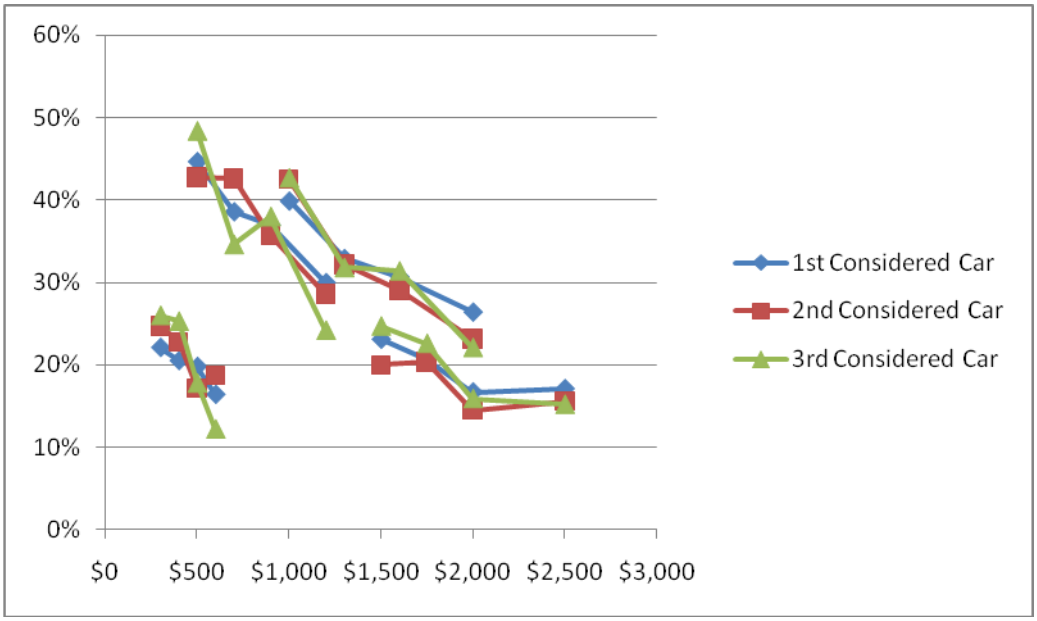


Figure 11

We also used counting analysis to examine cross effects for Exercise 2, finding many of them strongly significant.

Modeling via Multinomial Logit (MNL)

The Sawtooth Software community is quite familiar with three different utility estimation routines that all employ the logit rule: aggregate logit, latent class, and hierarchical Bayes (HB). We employed all three approaches for developing utility weights and building market simulators for these menu choice data.

Other approaches have been proposed in the literature, including multivariate probit (Liechty *et al.* 2001). We have not investigated multivariate probit. The purpose of this paper is to investigate how well the standard tools available to researchers can work in dealing with menu-based choice problems. In the future, we may find out that other methodologies also work well, or are even superior. Our hope is that the simpler models we propose here perform well enough to deliver accurate and robust results for the typical kinds of simulators managers demand. The excellent fit to holdouts we report below for this data set gives us hope that the MNL models may accomplish those aims.

We have experimented with three main approaches to specifying and coding the models for MNL estimation:

- Volumetric CBC Model
- Exhaustive Alternatives Model
- Serial Cross-Effects Model

You may think of these as tools you can mix-and-match to solve a variety of MBC problems. Again, these approaches just reflect different ways to code the choice sets. In all three cases, we are using MNL estimation.

VOLUMETRIC CBC MODEL

This model borrows a trick that we've been describing in our advanced CBC training workshops for some years now (and that is also described in a paper within these proceedings by Tom Eagle). The classic volumetric CBC example involves purchase of breakfast cereal. Imagine that eight product alternatives are available on the shelf, and the respondent is asked to state how many of each product she will purchase. This isn't a constant-sum task, as the respondent can allocate from 0 purchases to as many boxes of cereal as she thinks she can cart out of the store. Let's imagine the respondent completes 12 such tasks.

To model the data, we can simply scan the 12 tasks this respondent completed to identify the largest quantity of items ever purchased in a *single* choice task. Imagine that this maximum is found in task #4, where this respondent "purchased" the following quantities:

Purchase Volume: Task #4

Alternative	Quantity
Alternative 1	0
Alternative 2	3
Alternative 3	0
Alternative 4	2
Alternative 5	0
Alternative 6	0
Alternative 7	4
Alternative 8	0
Total:	9

The maximum volume for this respondent, for any *one* task, is 9.

To analyze the data, we'll generate a .CHS (or .CSV) file (the file for constant-sum allocation data supported by our CBC/HB and Latent Class software). A key thing to remember is that our CBC/HB and Latent Class systems automatically normalize the allocations within each choice task to sum to 100%. The trick, therefore, is to add a "None" alternative to the .CHS/.CSV file, so that the software believes that the respondent also faced a None alternative in each task. We reformat Task #4 as follows:

Purchase Volume: Task #4

Alternative	Quantity
Alternative 1	0
Alternative 2	3
Alternative 3	0
Alternative 4	2
Alternative 5	0
Alternative 6	0
Alternative 7	4
Alternative 8	0
None	0
Total:	9

Next, we also add a None alternative to the other 11 choice tasks. But, for these remaining tasks, the amount "purchased" of the None alternative is equal to 9 minus the volume purchased in that task. So, if in Task #1, the respondent "purchased" five boxes of cereal, a quantity of four will be given to the None. During market simulations, respondents are weighted by their maximum quantity value (9, for this respondent).

Now that we have introduced the coding of the model, it's very easy to see how to apply this to Exercise 1 for our automobile options study. We'll assume that the maximum number of items "purchased" for all respondents is eight (there are eight possible menu options).

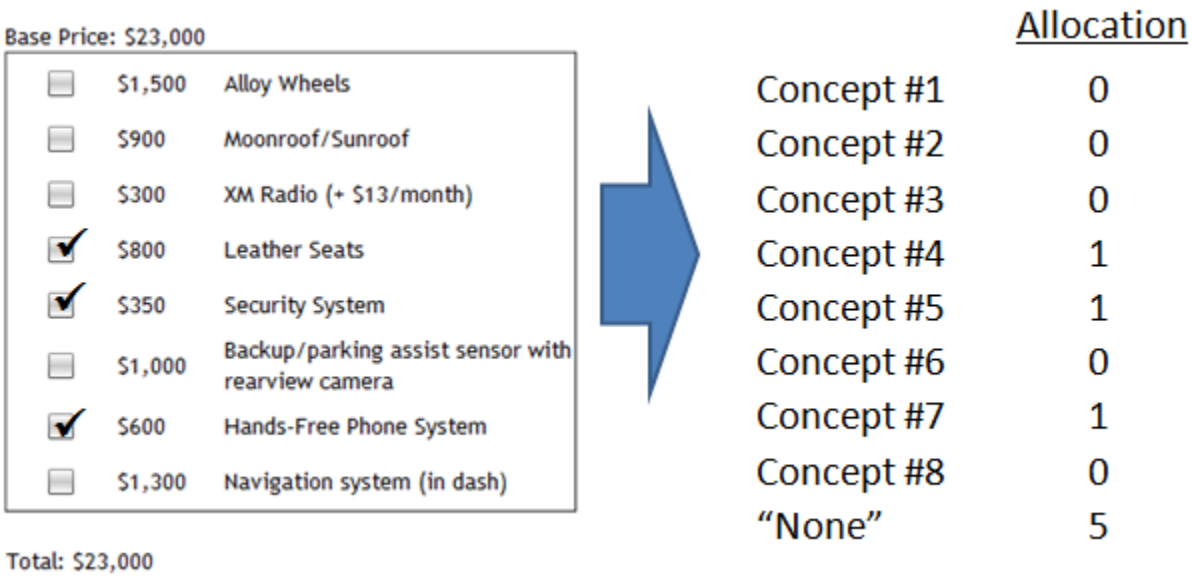


Figure 12

The next challenge is to code the independent variable matrix. To ensure reasonably fast convergence in our CBC/HB software (given its default settings for prior variance), we have found that it works best if the independent variables are coded to have absolute magnitudes in the single digits. It also tends to work out better if the independent variables are zero-centered. Our

preference[1] is to code the eight alternative-specific price variables, representing the eight options on the menu, as zero-centered with a range of 2. We also need to capture the inherent desirability of the eight options on the menu, plus the None alternative. We do this with K-1, or 7 columns coded either as effects-coding or dummy-coding, plus a separate column for the None parameter. Thus, the total number of columns in the independent variable matrix is:

7 dummy or effects codes (desirability of 8 menu options)

8 alternative-specific price coefficients (zero-centered)

1 dummy code (None weight)

= 16 total parameters

Once the CBC/HB model is run, in allocation mode, a set of utilities is made available in a .CSV file, and a simulator may be built in Excel. When using the logit rule to simulate choices for the sample for Exercise 1, these logit "shares of preference" need to be multiplied by 8 for each respondent (the weight representing the maximum number of items that can by "purchased" on the menu).

The key benefit of using the Volumetric CBC Model for MBC experiments is that it can estimate the likelihood of choosing multiple binary items on the menu using a single model. The main drawback is that it is ***not a theoretically sound model***, since the predictions are volumes rather than probabilities of choice, and those volumes are not bounded by 0 and 1.0.

Despite the obvious flaw in this model, it actually seems to work well for the data set employed in this paper. The aggregate predictions match the choice likelihoods for the holdout respondents very well. Aggregate predictions are one thing, but one may question what is happening at the individual level. For Exercise #1, even if we set all items on the menu to their lowest prices, 95% of the predicted "volumes" for alternatives (after multiplying the share of preference by a volume of 8) fell within the 0 to 1.0 range at the individual level. These, of course, are supposed to be likelihoods of choice bounded by 0 and 1.0, but the Volumetric CBC Model does not formally recognize these as likelihoods. It treats them as volumes of items purchased (without constraining the volume per person to be limited to 1). Even with all items set to lowest prices, the largest predicted volume for any one respondent (out of n=800) for any one item on the menu was 1.22. So, despite the theoretical shortcomings, the model seems reasonably well behaved.

EXHAUSTIVE ALTERNATIVES MODEL

This model assumes that respondents approached the menu task by considering *all* possible ways that the menu could be completed, and choosing the *one* most preferred way. For example, if the menu included just four binary items, there are 2^4=16 possible combinations of checks that can be done on the menu. To code the data for this example using the Exhaustive Alternatives Model, we treat each menu task as a discrete choice among 16 alternatives. Each alternative has a total price associated with it (or, the prices can be separated as item-specific price coefficients). The desirability of the 16 possible combinations is coded as K-1 or 15 dummy-coded columns in the independent variable matrix.

[1] We have described capturing the price effects as simple linear terms in this paper. Researchers should be on the lookout for non-linear price functions that are captured better using non-linear specifications, such as log-linear, quadratic, piecewise, or part-worth functions.

This approach only works well in practice when the total number of possible ways that respondents can complete the questionnaire is a reasonably small number, rather than in the multiple hundreds or thousands of possibilities.

For Exercise 1, there were $2^8 = 256$ possible ways that respondents could complete the menu. This would have resulted in each choice task being coded with at least 256 total columns in the independent variable matrix and 256 rows per task. With 800 respondents and 8 tasks each, such a problem becomes too large to run in reasonable time with HB estimation. We can reduce the size of the problem if we recognize that only about 150 of the possible 256 combinations were ever chosen by respondents (and assume the other combinations have zero likelihood of choice), but this still results in a very large problem.

But, for Exercise 2 (see Figure 5), there were only 48 unique possible ways to complete the menu ($3(2^4)$). This is quite manageable with CBC/HB software. Thus, each task may be coded as 48 alternatives, and the choice is coded as a single vote for one of those rows. The data matrix for one choice task is as follows:

48 rows Per task, 8 tasks per resp.

47 effects-coded columns	PriceBase	PriceOpt1	PriceOpt2	PriceOpt3	PriceOpt4	Choice
-1 -1 -1.....-1	-0.5	0	0	0	0	0
1 0 0..... 0	-0.5	0	0	0	-0.18	0
0 1 0..... 0	-0.5	0	0	0.17	0	1
...						
...						
...						
...						
...						
...						
...						
...						
...						
0 0 0..... 1	0.5	0.06	0.54	-.50	0.53	0

Figure 13

The inherent desirability of each of the 48 possible ways to complete the menu task is captured as a categorical variable with 48-1 = 47 effects-coded columns (row 1 in Figure 13 is the reference level). Next, we have captured separate price effects for the base price of the vehicle (PriceBase), and separate effects (PriceOpt1...4) for the prices for the options included in each way to complete the menu (note the zero prices when options are not a part of the alternative). Again, this reflects the idea that the respondent actually considered all 48 ways the menu could be completed, together with the price implications for each, and chose the *one* way with the perceived highest overall utility.

Once one estimates the model using MNL, there is some work on the back end to convert the predicted likelihoods for each of the 48 alternatives in the coded tasks into predictions of choices of items from the original menu. For example, option 1 on the menu might be coded in the "on" state only in alternatives 25 through 48. Therefore, one accumulates the total likelihood of respondents picking option 1 from the menu by summing the predicted likelihoods across coded alternatives 25 through 48.

The benefits of the exhaustive alternatives model is that it formally recognizes and predicts the *combinatorial outcomes* of menu choices, rather than just the marginal choices of each individual item on the menu. It is thus a more complete model of consumer choice. The drawback is that the model can become quite sparse at the individual level, with large numbers of independent variables and relatively few choice sets. This can lead to overfitting. Researchers may find that aggregate logit or latent class models do quite well with these exhaustive alternatives models, given the sparse nature of the data. Furthermore, because of the size of the models, runtime speed can become a real issue for HB.

Serial Cross Effects Model

This approach breaks the menu down into a series of separate choice models. For example, with Exercise 1 (Figure 3), we could treat this as eight separately run binary logit models. In each model, we are predicting the likelihood of selecting that menu item or not, given the inherent desirability of the item, its price, and the prices of other items on the menu.

Consider the choice of Alloy Wheels in Exercise 1. We may build a separate logit model to predict whether Alloy Wheels is selected or not (we treat "or not" the same as selecting the "None" within standard CBC model coding). The predictor variables include: the desirability of Alloy Wheels, the price of Alloy Wheels, and each of the prices of the other items in the menu. Conceptually, the model looks like Figure 14, where the arrows represent the effect of prices of different menu options on the choice of Alloy Wheels:

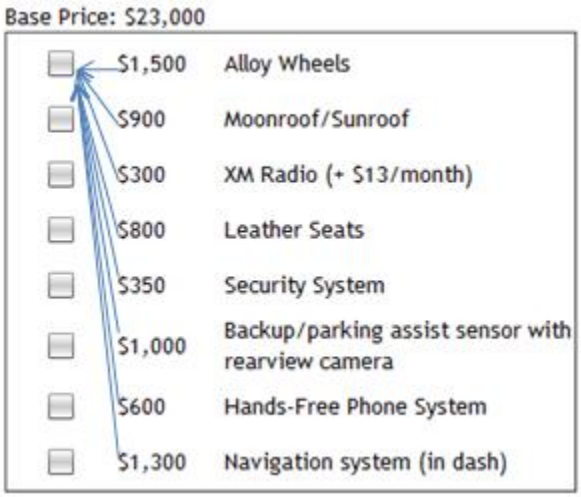


Figure 14

We build eight such separate logit models, where each model predicts the likelihood of picking a different item on the menu. The models are interconnected via the cross-effects terms (for example, the price of Moonroof affecting the purchase likelihood of Alloy Wheels, etc.).

While most menu studies in practice may use binary items (select or do not select), there are many menu situations that involve more than two mutually-exclusive choice outcomes, such as: 1) Standard cloth seats, 2) Black leather seats, 3) Cream leather seats. This is very simple to

manage with the Serial Cross-Effects Models and MNL. Rather than coding each choice set with two alternatives, we expand to incorporate three alternatives.

There is an old saying: The best way to eat an elephant is one bite at a time. The key advantage of the Serial Cross-Effect Model approach is that very complex menus may be broken up into smaller, digestible pieces. For each checkbox on the menu, we can develop a separate quite manageable model—especially if we only include the cross-effects that seem to have a significant effect on choice.

One of the main disadvantages of the Serial Cross-Effect Model approach is that it can be a hassle (and error-prone) to build so many separate models. Another problem is that if all possible cross-effects are allowed into the model, then the resulting what-if simulator may produce some strange results that are just due to random error. For example, decreasing the price of carrots on the menu may lead to a tiny (non-significant) *increase* in the likelihood for purchasing the fishburger. In reality, this effect may be non-significant, and if the relationship lacks face validity, it may only cause the client some consternation. Pruning the model of non-significant effects is one way to reduce this problem. Imposing utility constraints is another.

As with the Volumetric CBC Model, a big weakness with the Serial Cross Effects Model is that it doesn't formally recognize combinatorial outcomes (multiple items being selected together), but instead focuses on being able to predict the marginal choices of each separate item on the menu. While this may not be especially detrimental for predicting the average choice likelihood for items across the menu for the sample, it leads to less accurate individual-level predictions of the actual combinations of menu items selected. That said, most managers are interested in the average predictions of choice likelihood for the menu (given price changes) rather than predictions of the *actual combinations* that will be purchased. If the latter is the goal, then Cohen and Liechty recommend multivariate probit (Cohen and Liechty, 2007).

Some Menu Choices Depend on Other Menu Choices

A common hurdle to overcome in MBC questionnaires is when some choices on the menu may only be made if another choice has first been selected. For example, in Exercise 2 (Figure 5), the respondent cannot pick Alloy Wheels for vehicle #1 unless he has first chosen vehicle #1. For the fast food example in Figure 2, we restricted respondents from picking medium French Fries from the menu if they also picked a value meal (which included the Medium French Fries).

There is a straightforward way to handle dependent choices. Consider the choice of French Fries from the *a la carte* section of the menu in Figure 2. Respondents can only pick French Fries if they first rejected all value meals. Therefore, to predict the likelihood of respondents picking among the *a la carte* French Fry options, we only include in the model estimation the tasks where respondents rejected the Value Meals. Next, at the individual-level, we use the logit rule to predict the likelihood of choices on the menu. We multiply the likelihood that the respondent rejected all Value Meals by the likelihood from the second model (that involved task filtering) of picking each *a la carte* French Fry option. This formally recognizes via the choice simulator that respondents cannot pick both value meals and *a la carte* French fries.

For Exercise 2's automobile configuration task (Figure 5), we might similarly assume that respondents follow a 2-stage decision process: first, choose the vehicle; second, configure the chosen vehicle. To accomplish this, we build an MNL model that predicts the likelihood of

selecting vehicle 1, vehicle 2, or vehicle 3 (given their base prices and the prices of their options). This very much resembles a standard CBC MNL formulation, since it is a forced mutually-exclusive choice. Next, we can use either the Volumetric CBC Model, the Exhaustive Alternatives Model, or the Serial Cross-Effects Model, to predict the likelihood of selecting among the four options (Alloy Wheels, Moonroof, XM Radio, Navigation System), *given* that this vehicle (column) was selected. (Again, this is done by using only the choice tasks where the respondents picked the vehicle to develop the models that predict the selection of items within that vehicle.) When we simulate the likelihood of configuring options within each vehicle at the individual level, we multiply the likelihood of selecting that vehicle by the likelihood of configuring items on that vehicle (given the choice of that vehicle).

As we report the results below, we'll refer to this as the "2-Stage Model."

RESULTS

Earlier, we described that the data collection employed 1600 total respondents. Each respondent completed eight choice tasks, where the prices varied across the tasks. 800 respondents were randomly selected to be *calibration respondents*, used for estimating models and building a What-If simulator in Excel. These respondents were given one of 300 versions of the questionnaire, generated using CBC's design methodology. The other 800 respondents received an identical-looking questionnaire, except that these respondents completed one of just 3 versions of the questionnaire, again generated using CBC's design methodologies. Thus, on average, each of the three holdout questionnaire versions was answered by 800/3 = 267 respondents.

We simply tallied the percent of holdout respondents who chose each of the items, for each of the 8 choice tasks x 3 questionnaire versions = 24 menu tasks. With Exercise 1, there were 8 items that could be checked on the menu. Thus, there were 24 tasks x 8 menu items = 192 separate holdout probabilities to be predicted using the market simulator. This reflects a great deal of holdout information for validating the models.

Exercise #1 Results

The R-Squared fit for holdout predictions and the Mean Absolute Error (MAE) of prediction, under the Volumetric CBC Model, were as follows:

	R-Squared	MAE
Volumetric CBC/HB model	0.925	0.0370
Volumetric CBC/HB model (with covariates)	0.928	0.0358

Scatter Plot: Predictions vs. Actual Holdout Probabilities Using Best Model (with Covariates)

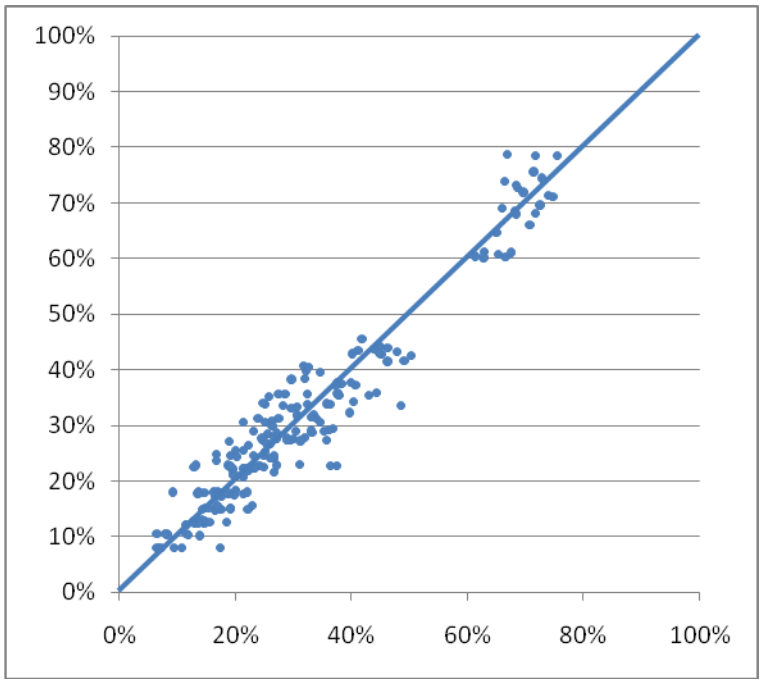


Figure 15

These predictions look very good, in the aggregate, especially considering that we cannot expect to achieve perfect predictions due to sampling error. Predictions based on 800 respondents were compared to actual choices for *different choice scenarios* completed by a *different group* of 267 respondents. Even if respondents answered without error, and our model specifications were perfect, there would still be unexplained error due to sampling error (a group of respondents predicting a separate group of respondents).

Because we found no evidence of significant cross-effects, we did not attempt the Serial Cross-Effects Model. Because there are 256 possible combinations of selections from this menu (resulting in huge data matrices and glacial run times), we did not attempt the Exhaustive Alternatives Approach. We focus much more attention on Exercise 2, which was a more complex and interesting data set.

Exercise #2 Results

The predictive fit to holdout choices for Exercise 2 was as follows, for the different approaches:

	R-Squared	MAE
2-Stage model (Aggregate Logit)	0.965	0.0201
2-Stage model (Latent Class)	0.960	0.0223
2-Stage model (CBC/HB)	0.960	0.0225
2-Stage model (CBC/HB, with covariates)	0.961	0.0224
Exhaustive alternatives model (Aggregate Logit)	0.954	0.0231
Exhaustive alternatives model (Latent Class)	0.956	0.0229
Exhaustive alternatives model (CBC/HB)	0.956	0.0234
Exhaustive alternatives (CBC/HB, with covariates)	0.957	0.0229
Serial cross-effects model (Aggregate Logit)	0.954	0.0226
Serial cross-effects model (Latent Class)	0.952	0.0236
Serial cross-effects model (CBC/HB)	0.942	0.0265
Serial cross-effects model (CBC/HB, with covariates)	0.951	0.0249

Scatter Plot: Predictions vs. Actual Holdout Probabilities Using Best Model (2-Stage Aggregate Logit)

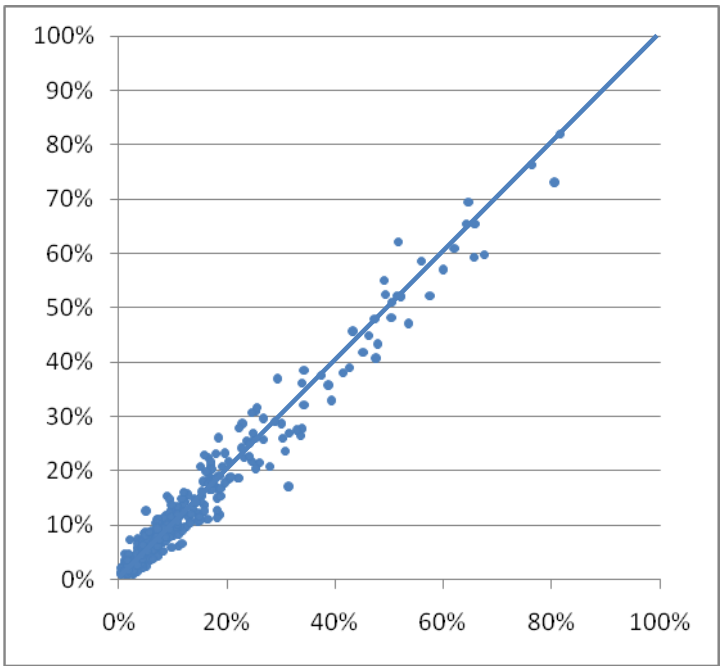


Figure 16

In general, all attempted models worked well, with R-squared values better than 0.94. Given that we cannot achieve perfect fit to holdout data given sampling error, we are doing about as well as possible.

For Exercise 2, we also tried aggregate logit and latent class. It is very interesting to note, and somewhat surprising, that the aggregate logit approach achieved the highest predictive validity (the 2-Stage model, with R-squared of 0.965) for this data set. This is surprising, since our experience with traditional CBC is that HB generally shows higher predictive validity for holdouts than aggregate logit (at least generic aggregate logit models *without* cross-effects). Aggregate logit is notoriously prone to IIA difficulties, and one of the benefits of HB is the ability to reduce IIA troubles for standard CBC studies. But, with MBC models, IIA should be less of a concern. IIA is concerned with maintaining the ratio of choice likelihoods for competing alternatives, when a given alternative is added or modified within a choice scenario involving at least three alternatives. If using a series of binary logits to estimate choice likelihoods of each item on the menu, there are only two alternatives in each model. Furthermore, differential substitution effects among items can be accounted for using cross-effect terms (which we've done here with the aggregate models). As a final point, IIA is less of a concern if the full context of all available products was shown in each choice task, which is often the case with MBC studies[2].

[2] MBC studies may also vary the presence of items on the menu (availability designs). When this is the case, the availability of an item may be used as a cross-effect for predicting the likelihood of choosing other items on the menu. With availability designs, HB models may have an additional advantage over aggregate models, but this remains to be proven with additional datasets.

Given the amount of evidence across the industry in favor of HB methods, it would be unwise to dismiss HB modeling as unnecessary for MBC studies based on our findings with this one dataset. HB provides the liberating convenience of easy on-the-fly filtering by respondent segments in the market simulator, without having to go back to square one and re-estimate the model (as with aggregate logit) for each segment. This benefit alone will make it worth the effort for most practitioners to use HB rather than aggregate logit. Also, the only standard for success reported here is aggregate predictive validity. We haven't worried about individual-level prediction, especially prediction of combinations of choices (such as reported in Figure 8). Individual-level models would seem to have an upper hand if this were the goal.

The approaches that formally recognize that there were logical exclusions in the way that respondents could complete the menu (prohibited choice combinations) tended to perform better than the method (Serial Cross-Effects Model) that ignored these exclusions. This seems quite logical and reasonable, and researchers should take care to use models that formally recognize any logical exclusions within MBC questionnaires. If we were modeling a questionnaire that didn't include any logical exclusions, the Serial Cross Effects model may have been the preferred model. For example, at this conference, Chris Moore of GfK reported solid results for a restaurant menu study when using the Serial Cross Effects approach.

As a final note, using covariates within CBC/HB makes very little difference in predictive accuracy over standard HB, though the results suggest perhaps a tiny directional improvement. This is in line with other research we've conducted on covariates, and the results of two other papers delivered at this conference (Sentis; Kurz and Binner). Based on our experiences with covariates in HB, their real value would be seen in enhanced discrimination among market segments if reporting MBC what-if simulations by segment.

ANALYSING PICK N' MIX MENUS VIA CHOICE MODELS TO OPTIMISE THE CLIENT PORTFOLIO – THEORY & CASE STUDY

CHRIS MOORE
GFK NOP

BACKGROUND

The way consumers shop for products or services has changed dramatically since the evolution of the web and as a result of technological advancements and the high penetration of the web in advanced countries, businesses have started to move away from offering consumers fixed products or bundles of products. This emphasis on mass customisation has led to the development of menu-based systems. That is, where the previous emphasis was based on designing a portfolio of 'fixed' products from which consumers had to choose from, irrespective of whether it fitted their needs, the emphasis has changed such that the individual features of the product are presented in a series of menus and are individually priced allowing consumers to tailor their ideal product configuration within a specified budget.

In addition to individually priced features these menu systems typically allow consumers to choose from a number of pre-defined products which generally offer a discount over buying the individual features that are contained within the product. The reason for this switch to mass customisation of products is to entice purchasing by meeting the consumers unique needs for the product or service.

This change in the way businesses now market their products has meant that traditional modelling approaches such as conjoint analysis are less appropriate and new innovative ways to analyse data are needed. Cohen and Liechty (2007) commented that:

"One might think that traditional conjoint analysis is appropriate and effective for understanding how people want to construct product bundles, but the menu situation's additional complexity makes that approach entirely inadequate. Enter menu-based conjoint analysis – expressly designed for handling a build-your-own, select-from-a-menu, mass customisation situation"

A case study will be presented that details the findings from a major study conducted by GfK NOP in partnership with TGI Friday's where the objective was to develop a menu-based optimisation tool that will allow the identification of key items from their menu and allow TGI Friday's to optimise the pricing of these key items in order to maximise net profit.

We will first discuss the merits of using traditional conjoint analysis as a technique to answer this research problem then look in detail at one of the approaches to show how we can analyse menu-based choice systems and produce dynamic multi-menu optimisation tools.

PRINCIPLES OF CONJOINT

Conjoint analysis is concerned with understanding how people make choices so that businesses can design new products and/or services that better meet consumers underlying needs. It is a popular Market Research technique and has been found to be an extremely powerful of

way of capturing what really drives customers to buy one product over another and what customers really value. Rather than asking respondents directly what they want in a product, conjoint techniques employ a more realistic context for respondents to evaluate potential product profiles.

A key benefit of conjoint analysis is the ability to produce dynamic market models that enable businesses to test what steps they would need to take to improve their acquisition and/or reduce churn, based on potential competitor behaviour.

Conjoint analysis works by decomposing a product in terms of attributes and levels. An attribute is a general feature of a product e.g. Brand, Engine size, Transmission, Price etc. Each attribute is then comprised of a number of specific levels. So, for the attribute 'Transmission' we might have the levels: Manual, Semi-Automatic and Automatic. Levels are alternatives that are feasible for the business to provide and need to be understood in order to determine the most optimum product.

Modelling techniques are used to produce 'Part-worths' (also commonly referred to as Utilities) for each level of every attribute tested. The part-worths are a measure of the value or attractiveness of each attribute level to the respondent. The higher the part-worth the more 'worth' a respondent has put on that level and therefore how much more desirable the level is over other levels within the same attribute. A measure of importance can also be derived for each attribute, which is based on the range of part-worths within each attribute and compared against the range of part-worths across all attributes.

Choice Based Conjoint vs. Menu-based Choice Systems

Traditional conjoint analysis has its origins dating back to the early 1970's and with the invention and ever-increasing capabilities of the PC it can now deal with extremely complex research problems. There are many types of conjoint techniques available to the analyst, each of which has its own advantages and disadvantages.

Since the late 1990's with the use of Hierarchical Bayes (HB) estimation becoming commercially available, Choice Based Conjoint (CBC) is currently the most widely used conjoint technique (Sawtooth Software Customer survey, 2010). Estimation via Hierarchical Bayes has been described very favourably due to its ability to provide robust individual level part-worths given only a small amount of information. It does this by 'borrowing' information about the population (respondent means and co-variances) to aid creating part-worths for each individual (Orme, 2000). Prior to HB becoming available, early methods for analysing CBC data involved pooling data across respondents and analysing choice behaviour through an aggregate Multinomial Logit model (MNL). This often masked important findings especially when analysing attributes that are categorical in nature and/or where the population is heterogeneous. A further disadvantage of the aggregate MNL model is that it suffers from the IIA assumption (Independence from Irrelevant Alternatives). This implies that the choice ratio between two products does not change by the addition of new products (Sawtooth Software, 2009). As a result, more traditional conjoint methods such as traditional card sort (CVA) were often favoured as the data collection for each respondent provides richer information, which enabled the analyst to calculate part-worths at the individual respondent level.

CBC works by showing respondents a number of pre-designed product concepts, typically 3-5 at a time, known as a Task, and asking which concept or concepts are preferred. These tasks are repeated a number of times with the product concepts changing each time.

The main advantages of CBC over other conjoint techniques are that it is perceived to be a simpler task for respondents to complete, namely, it mimics buying behaviour more realistically as consumers will be choosing from a small sub-set of products and picking the most suitable product, rather than ranking or rating many products. Typically, there are fewer tasks that are needed to be shown for the same type of conjoint design as it is common practice to produce many versions of the tasks enabling the respondent to only have to evaluate a small number of these tasks.

In addition to main effects, higher order utilities can be calculated to take into account that two or more attributes may interact. For example, if the combination of 'Red' and 'Ferrari' results in a greater preference than the main effect part-worths suggest then interaction terms can be employed to model this additional effect. A further advantage of CBC is the ability to be able to include a 'None' option as part of the task as well as in simulations. Again, it is perceived to better mimic real world scenarios since respondents should not have to select from concepts where none of them are desirable. It can be further used as the constant alternative to reflect their current situation e.g. I prefer to stay with my current supplier.

However, despite the rapid growth in popularity there are several issues to the Choice Based Conjoint technique that need to be considered when we think about analysing menu-based choice systems:

1. Designs are generally limited to **no more than 6-7 attributes**; respondents can only be expected to have a limited amount of enthusiasm for any trade-off task and if they perceive the series of trade-off questions to be too long this is likely to lead to data of questionable quality. The problem with having too many attributes is that respondents will not consider all of them, rather focussing on only the few that are most salient to them leading to obtaining poor quality information.

2. In order to create efficient designs that are (near) orthogonal a proportion of **tasks shown to respondents tend to be unrealistic** to both the client and the consumer. If the concepts shown to consumers are not close to their ideal product then this can create the perception that the survey is not focused or relevant to them. Further to this it does not allow respondents the opportunity to select a product that suits their needs and limits them to choosing a fixed product from a subset of fixed products.

3. **Some respondents display non-compensatory rules** (Gilbride et al, 2004 and Hauser et al, 2006) - that is, there may be levels from a particular attribute that have to be present in order for the respondent to consider the product. Design strategies such as Sawtooth Software's Complete Enumeration heavily favour minimal overlap, meaning that when the number of levels in an attribute is equal to or greater than the number of concepts that are shown in a task a level will generally only be present in one of the concepts. For respondents who display non-compensatory rules this only leaves them with the option of choosing this single concept or the None option (or not having an option at all if the level is not present) and therefore losing valuable information about their preferences. It should be noted that strategies such as Sawtooth Software's Balanced Overlap can overcome some of these issues.

4. In product areas where multiple products can be chosen in a task there may be a need to determine **potential changes to the size of the market** in order to understand metrics such as profit. Standard conjoint analysis cannot deal with this appropriately therefore there is a need to create models that can determine changes in market size effectively. A Volumetric CBC approach described by Sawtooth Software in their literature goes some way to answering this problem when there are no limitations between choices within a task.

5. In standard conjoint designs the levels of **attributes are independent of the total price** of the product as a whole. This leaves situations where you have concepts that are feature rich yet are shown together with a low price level or vice versa. Strategies such as conditional pricing and alternative-specific designs can overcome some of these issues but can quickly become too complex when multiple attributes are involved.

Recent developments such as Adaptive Choice Based Conjoint (Johnson, Orme 2007) have addressed a number of these issues but with product offerings becoming more diverse and increasingly complicated in terms of offering bundles and/or individual menu style features there is a need to be able to find a solution to optimising the pricing of individual features across multiple menus to increase profitability for the business.

Wind and Mahajan (1997) recognized the importance of researching mass customised products and commented that:

"From a marketing research point of view, the focus is no longer on conjoint analysis studies leading to the identification of an optimal product or product line, but rather on the following:

1. The identification of the set of features and levels that typically constitute the conjoint analysis tasks.

2. The way consumers want to customise their products.

3. The premium, if any, customers are willing to pay for a customised design verses an off-the-shelf product.

Hence, the key marketing challenges are to (1) identify the features that should be offered; (2) understand how customers want to build customised products and which customers want to customise and which do not; and (3) uncover how to price each feature to simultaneously increase customer value, revenues and profits."

Most literature in this area refers to analysis of this type of problem as Menu-based Choice analysis. The basic premise of how CBC tasks are presented can be modified such that they are presented in a series of menus from which the respondent can choose individual features from one or more menus in order to build up their desired product. As a result of this multi-choice process, within menu-based choice systems there may be a need to be able to simulate shares that take into account changes in market size, in order to gauge the number of individual features that respondents are purchasing and therefore enable the analyst to calculate revenue/profit figures. Further, menu-based choice systems are likely to engage respondents more as you are presenting realistic choice options rather than fixed products that are inflexible and constrained to a fixed price. There is also the option in menu-based systems to allow respondents to choose between fully customised products and off-the-shelf pre-bundled products that are cheaper than the sum of the prices of the individual features within the bundle.

The issue of non-compensatory behaviour is very much reduced in menu-based choice systems as all features are offered to the respondent all of the time, though it still may not be able to deal with consumers that will only consider a product below a certain price point. There is also fewer limitations on the number of attributes that can be accommodated as attributes can be split up as individual menus so the cognitive burden on respondents is reduced and there is not the need as in CBC for the respondent to have to make extremely complex choice decisions, which typically involves simultaneously processing several products with many levels and determining what are important to them and how they will trade off levels across the different products.

PREVIOUS RESEARCH

The use of product configuration is not a new area of analytics and simplistic analysis has long been available such as subjecting respondents to Build Your Own (BYO) tasks. Within these tasks respondents indicate which features/levels they would select for a number of attributes, where each level is commonly priced to provide a penalty effect of choosing the more desirable levels from each attribute. However, price elasticities are typically in the form of counts, which would only allow the client to identify potential cannibalisation to/from an item based on a single item changing price. There are numerous drawbacks to this technique, not least that respondents want to upgrade from the lowest levels of attributes (Johnson et al, 2006).

Of more significance, there has been a number of menu-based choice modelling approaches described in previous literature including: Liechty, Ramaswamy & Cohen (2001), Bakken & Bond (2004), Cohen & Liechty (2007) and others.

Cohen & Liechty (2007) compared three approaches to modelling menu-based systems which are:

a. Analysing data as a **series of binary choices** without regard to combinations – The menu choices are converted into a series of K binary choices (where K is the number of features in a menu). The modelling then estimates each binary choice separately. As long as menu choices are uncorrelated and there are no constraints between the individual menu choices then this approach can work well but when menu choices are positively correlated the estimates obtained will be inaccurate.

b. **Traditional choice modelling** – The menu choices are converted to a single choice array from 2^k possible arrays and a multinomial choice model is used to analyse the chosen array as a single choice from all possible arrays. While this gives the ability to evaluate pre-bundled options as the number of features increase the size of the exploded choice set becomes excessively large.

c. **Menu modelling** – The individual menu choices are preserved and the analysis models choices in terms of the probability of choosing a collection of features. A utility is specified for each feature and if the utility is above an estimated threshold then it is chosen. The modelling can provide an evaluation of pre-determined bundles as well as reveal 'natural' bundles and further yields correlations between the errors in estimating the utilities of features, net of price and other effects.

Similar to model (a), Orme (2007) described a menu-based choice methodology using a fast food restaurant example. Using an online methodology with 681 respondents, sixteen different items from a menu were tested within a Choice experiment, with each of the items being tested at

four price points. Bundled meal deals were included in the experiment and tested with four levels of discounting, with the price of the bundle being based on the price of the individual items within the bundle minus the discount. Respondents evaluated 8 choice tasks, where in each choice task the price of the items varied and respondents were asked what item(s) they would choose from the menu. Respondents could either choose from one of the bundled meals or choose multiple items from the á la Carte menu. They were additionally offered the option to not buy anything from the menu and go to a different outlet or eat at home.

Analysis of the data comprised of splitting the design in to its category menu parts, namely: Meal deals, Sandwiches, Fries, Drinks, Salads, Healthy sides and Desserts. Each of the seven categories from the overall menu were analysed as separate choice models using Hierarchical Bayes estimation to determine preference for each of the items within each category menu. The seven models were then combined through a logit rule to produce an overall model.

For each component part of the menu, the probability of choosing an item within that component menu is a function of:

$$Prob(d_i) = f\,(Desirability d_i + Price d_i + Price a_1 + Price a_2 + Price a_3 + Price a_4 + Price a_5 + Price a_6 + Price a_7 + Price a_8 + Price a_9 + Price a_{10} \ldots\ldots)$$

Where:

$Prob(d_i)$ = Probability of choosing item i

$Desirability d_i$ = Desirability of item i

$Price d_i$ = Price of each item alternative in the category

$Price a_1$... $Price a_n$ = Price of items not in category

Due to the number of parameters to be estimated, Linear terms were modelled for each attribute, which resulted in 119 total terms to be estimated (7 models * 16 (items) + 1 (None)). The resulting 7 choice models were then linked together through an excel-based logit rule simulator to produce a single model that contained full main effects and cross-effect terms.

Two holdout tasks were incorporated into the conjoint experiment and yielded Mean Absolute Errors (MAE) of 1.6% and 1.3% across all items with a maximum deviation for any single item of 4.4% and 5.3% respectively. Further analysis showed price elasticity comparisons between the counts analysis and the simulated data and results showed only minor differences. Orme cautioned that as a result of the number of cross-effects there will be some generally non-significant reversals in preference shares due to random noise, such as increasing the price of small fries that leads to a decrease in the demand for ice cream. In the 2010 Sawtooth Conference, Orme described this approach as the 'Serial cross-effects' model.

Eagle (2009) presented a further solution that used a joint discrete continuous approach. In step 1, choice models are created to capture the patterns of substitution while in step 2 a regression-based approach is used which takes predictions from the choice model as independent variables.

The approach used in this case study follows the methodology that Orme presented so all further references will relate to this work.

CASE STUDY

GfK NOP was commissioned by TGI Friday's to develop an optimisation tool that would allow them to optimise the pricing of key items on their menu to maximise profit. In addition to individual items, the menu comprised of 'Value' meal deals which bundle together multiple courses and offer a significant discount in price compared to purchasing the individual courses separately.

A key consequence of changing the price of these key items is the potential cannibalisation to and from key competitors so analysis would need to result in capturing likely changes in market share, in terms of the number of covers to and from four main competitors. The analysis needed to combine the results of the competitor analysis together with the optimum configuration of priced items in order to be able to calculate profit.

A 2-stage Choice approach was conducted to mimic consumer habits when choosing to eat out at a restaurant.

Stage 1 – Comparing TGI Friday's menu against competitor menus

This stage was designed to evaluate the impact of changing price on key items from the TGI Friday's menu on outlet choice against 4 key competitors. Respondents evaluated the TGI Friday's and a competitor menu, side by side, with only the prices of items in the TGI Friday's menu varying each time. Respondents with multiple competitors in their area were first asked to choose between fixed priced competitor menus to determine their preferred menu. The TGI Friday's menu was then offered at varying prices for each item against the preferred competitor and respondents were asked to select which menu they would choose, from the two available to them. At this stage respondents were not offered the option to eat out at a different restaurant and it was assumed that they would eat at one of the 5 restaurants in the research design. Analysis of this data was conducted using a standard CBC approach where the competitors were included in the analysis as fixed alternatives.

Stage 2 – The relative price of items within the TGI Friday's menu

This stage was designed to evaluate the choice and price sensitivity of the key items within the TGI Friday's menu. Respondents evaluated a number of menus with the prices of the items varying each time. Respondents were asked to select which item(s) they would choose from those available and they had the option to leave the restaurant if none of the items presented to them at the price indicated were appealing. Analysis of this data was conducted using the Serial cross-effect model previously described.

Research Design

The sample design consisted of an online interviewing methodology using GfK NOP's online panel, with 1,490 respondents (1,602 before cleaning) that are representative of people aged 18 – 35 (Young Adult or Family life stage), who eat out monthly or more often, in 'branded restaurants serving alcoholic drinks and offering full table service'. All respondents had to have a TGI Friday's in their area (area was unspecified and subjective for each respondent), though not necessarily had to have been to a TGI Friday's, have at least one of four key competitors in their area and have eaten at one of these 5 restaurants in the last month, the most recent occasion being 'the occasion of relevance' for the survey.

Following U&A demographic and screening questions, the questionnaire collected key detail about 'the occasion of relevance'. This included the type of occasion (Party, Birthday etc), who they ate with, the time of day and the day of the week. Thereafter, the questionnaire focused on price sensitivity through choices between and within menu scenarios based on this most recent occasion.

In each stage, 12 choice tasks were shown to respondents, of which two were holdout tasks.

A 20 attribute (Item) design was generated (where each attribute is an item from the menu), with each item having 5 price levels. The items were split out across the different menu categories and comprised of 5 Starters, 2 Value deal bundles, 10 Main courses, 2 Desserts and 1 Drink option. In order to provide a more comprehensive menu for respondents to evaluate, multiple desserts were shown to respondents but were collapsed at the analysis stage as they were shown at the same price point in each choice task. Four other types of drinks were also offered on the menu at fixed prices but were subsequently not analysed in the simulation model.

The items tested represented TGI Friday's main volume items and accounted for the following volumes:

Starters:	53.3%
Value deals:	100.0%
Mains:	42.4%
Desserts:	62.9%

Both stages were put in the context of the respondents most recent occasion and they had to decide what they would personally choose, that is, at Stage 1, which restaurant they personally would prefer to eat at and in Stage 2, what they would personally choose to eat themselves once in the restaurant. It should be noted that two starters and two desserts were 'sharing' dishes. The text used in Stage 1 and Stage 2 is detailed in figure 1 and figure 2.

A design limitation was imposed such that respondents could only select one option from each of the category menus, that is, respondents could not select two starters for example. Analysis of standard survey data indicated that 97% of respondents chose a main course so it was pragmatic to assume that this limitation would have no inherent bias on the final results.

TGI Friday's offer Value meal options on a Monday-Thursday, which offer the consumer the option to purchase two or three courses for a significant reduction in price compared to purchasing the courses individually. Consumers can only choose from a selected number of items and this was reflected in the survey by a number of items having an asterisk next to them. When the most recent occasion for a respondent was on a Friday – Sunday these value meal options were removed from the menu in order to produce a more realistic scenario. Further, if a value meal was chosen in the choice experiment it was not possible to choose any other food items from the menu and vice versa.

Equivalent items were used for the competitor menus where possible and all menus were branded in the house style to avoid any bias.

Figure 1

Still thinking about this last occasion, dinner with your partner, on a Thursday Evening, on the next few screens you will see the menus from two restaurants with a number of dishes listed at different prices.

At each screen please select the restaurant that you personally would prefer to visit for this occasion based on the menus shown.

The two menus will be repeated several times and each time the prices on one of the menus will change. Please base your decision on the menus presented to you at the prices shown.

Figure 2

Please still think about this last occasion, dinner with your partner, on a Thursday Evening.

The next screen shows just one restaurant menu. We would like you please to select what you personally would choose to eat and drink from this menu.

The most that you can choose is one option from each section:

- **Starter**
- **Main**
- **Dessert**
- **Drink**

It is not necessary to choose an option from every section - if you would not choose a starter or dessert, leave this section blank. The menu will be repeated several times at different prices. For a fuller description of each of the menu items, please hover your mouse cursor over the desired choice.

If you would not eat anything in this restaurant then there is the option to leave the restaurant.

HOLDOUT TASK VALIDATION

When assessing the validity of choice models it is typical to include holdout tasks in the survey. Holdout tasks are held out from the analysis and a number of validation statistics can be calculated to compare the simulated results verses the actual results obtained from the holdout task:

Mean Absolute Error (MAE)

A common measure defined by the average of the absolute errors between the simulated choice probability and the actual choice probability.

Percent Absolute Error (PAE)

If there are lots of items in the choice experiment then the expected choice probability goes down and subsequently MAEs also will go down. Therefore MAE's can give a false sense of

accuracy when there are many items and it is also not possible to compare MAE's across studies when the number of items is different.

The PAE is calculated by dividing the MAE figure by the average actual choice probability across the items in the holdout task. MAE is then relative to the expected share.

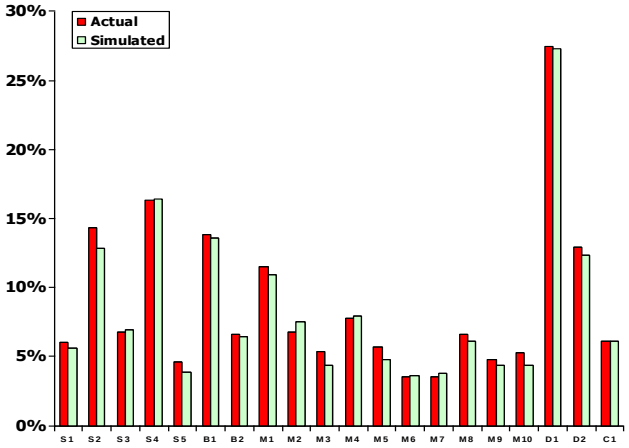
R^2

A plot of the actual vs. simulated choice probabilities is generated and a linear line of best fit is created. The R^2 shows proportion of variability that is accounted for by line of best fit

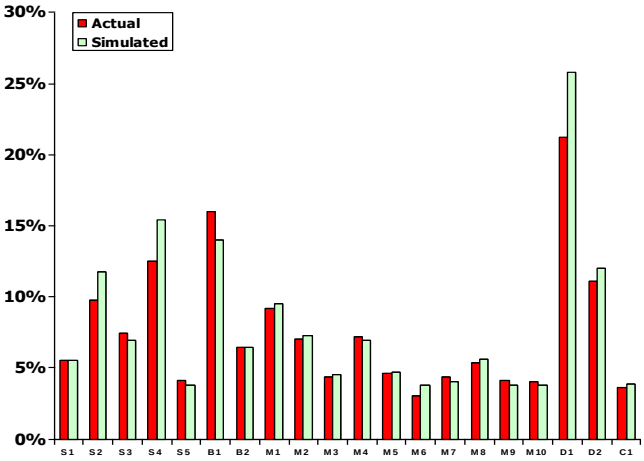
Figure 3 shows the comparisons of the holdout results and the simulated results for the menu-based choice stage. In addition to the actual differences, the MAE, PAE and R^2 have been reported for both holdouts. For both holdouts, the validation measures are extremely robust.

Figure 3

Holdout 1



Holdout 2



Note: Actual choice probability is the left bar

	Holdout 1	Holdout 2
Mean Absolute Error (MAE)	0.48%	0.82%
Percentage Absolute Error (PAE)	5.40%	10.87%
R^2	0.992	0.956

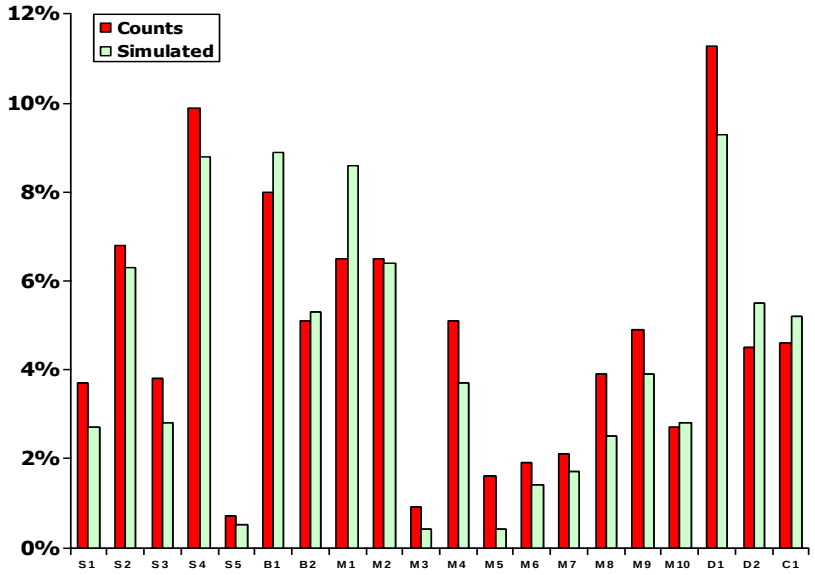
SENSITIVITY ANALYSIS (OWN EFFECT)

Figure 4 shows the sensitivity of each item as its own price changes from the lowest price point to the highest price point. All other items are set to the current TGI Friday's base price. Comparisons were made between the raw counts and simulated data to identify whether the

range of elasticity has been captured in the model as well as identifying the accuracy of the individual price points. Overall the simulated shares at each price point are in line with the raw counts and the range of elasticity has been captured in most items. In the simulated model, in 14 of the 20 items, sensitivity was slightly lower than raw counts but the model was not adjusted for scale in order to account for this.

Figure 4

Range of Elasticity



Note: Actual choice probability is the left bar

	Range of Elasticity	Individual Elasticity
Mean Absolute Error (MAE)	0.86%	0.68%
Percentage Absolute Error (PAE)	18.15%	8.08%
R^2	0.891	0.977

SENSITIVITY ANALYSIS (CROSS EFFECT)

Figure 5 shows the sensitivity of each item as the other items change price from the lowest price point to the highest price point. All other items are set to the current TGI Friday's base price. As one would expect, the elasticity within category menu is positive, for example, when the price of Starter 2 increases then the share of Starter 1 will increase (by 0.4%). It is interesting to note that there are a number of negative cross effects but this is not unexpected. For example, by increasing the price of a main course to the highest price point, due to budget constraints the customer may decide not to have a starter and/or dessert and therefore the share for the

starter/dessert decreases. Similarly, a positive cross-effect can be achieved when increasing the price of a main dish may make the customer switch to a cheaper main course and therefore allow them to have a starter/dessert and still remain within the same budget.

It is interesting to note that other than Main course 1 (which was the cheapest main dish) there is generally a low elasticity between the main courses suggesting that rather than changing between main courses when price increases, customers are choosing to change between starters/desserts or not select a starter/dessert at all.

Figure 5

		Effect on dish																			
		S1	S2	S3	S4	S5	VM1	VM2	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	D1	D2	DR1
Changing price of dish from low to high price	S1		0.7	0.1	1.1	0.0	-0.5	-0.3	-0.6	0.0	-0.1	0.0	-0.1	0.1	0.0	-0.1	-0.2	0.0	0.1	-0.4	-0.1
	S2	0.4		1.0	2.8	0.1	-0.9	-0.5	-0.8	-0.4	0.1	-0.3	0.1	0.1	0.0	0.1	0.1	0.0	-0.4	0.2	0.8
	S3	0.1	1.2		0.9	0.0	0.5	0.7	-0.6	-0.1	-0.2	-0.1	0.1	0.1	0.0	-0.2	-0.3	-0.2	0.0	0.1	-0.2
	S4	0.7	0.7	0.8		0.2	-0.4	0.3	-0.7	-0.4	0.1	-0.4	-0.1	-0.1	-0.2	-0.1	-0.2	-0.2	-0.2	-0.2	-0.1
	S5	0.0	0.1	0.0	0.2		0.0	-0.3	-0.2	0.0	0.1	0.0	-0.2	0.0	-0.1	0.2	0.0	-0.3	-0.5	0.7	1.1
	VM1	0.3	0.5	0.2	0.4	0.5		2.9	1.3	0.1	0.1	0.5	0.3	0.1	0.0	0.3	0.5	0.4	0.8	0.6	0.5
	VM2	0.0	0.3	0.1	0.1	-0.1	4.1		-0.1	0.0	0.0	0.2	-0.1	-0.1	0.0	-0.1	0.0	0.1	-0.5	-0.1	-0.6
	M1	-0.2	-0.8	-0.1	-0.5	0.0	1.9	0.3		1.8	0.2	1.4	0.1	0.1	0.1	0.3	0.8	0.9	-0.9	-0.9	0.2
	M2	-0.1	-0.1	-0.1	0.1	0.0	0.2	0.2	2.2		0.1	0.9	0.1	0.2	0.1	0.5	0.5	0.8	-0.8	-0.5	0.5
	M3	-0.1	-0.3	0.0	0.4	0.0	0.0	-0.2	0.1	0.0		0.1	0.1	0.0	0.0	0.1	0.0	0.0	1.0	0.3	0.3
	M4	0.1	-0.3	-0.2	0.1	0.0	0.2	0.3	0.9	0.5	0.3		0.2	0.1	0.1	0.2	0.3	0.4	0.1	0.0	-0.4
	M5	0.2	0.4	0.1	-0.1	0.1	0.1	-0.2	0.0	0.0	0.1	0.2		0.0	0.0	0.1	0.0	0.0	0.7	-0.1	0.1
	M6	0.2	0.1	0.0	0.4	0.0	-0.4	-0.1	0.1	0.1	0.0	0.3	0.1		0.2	0.2	0.2	0.0	-0.4	-0.3	0.0
	M7	-0.1	-0.6	0.0	-0.6	0.2	-0.9	-0.5	0.1	0.2	0.0	0.2	0.0	0.2		0.4	0.2	0.0	0.4	-0.1	-0.8
	M8	0.0	0.4	0.1	0.1	0.0	-0.2	0.1	0.2	0.4	0.0	0.2	0.1	0.1	0.3		0.2	0.2	0.0	0.0	0.8
	M9	0.1	-0.4	-0.1	-0.9	0.0	1.6	0.0	0.8	0.4	0.0	0.5	0.1	0.1	0.2	0.4		0.7	-1.4	-0.9	-0.3
	M10	0.3	0.3	0.1	1.1	0.4	0.2	-0.5	0.7	0.5	0.0	0.4	0.0	0.1	0.0	0.1	0.5		0.9	0.9	1.2
	D1	-0.1	-0.3	0.0	-0.4	0.0	0.6	0.4	1.0	-0.5	-0.2	-0.4	-0.1	-0.1	0.0	0.0	-0.4	-0.3		3.0	0.0
	D2	0.1	0.4	0.2	0.7	-0.2	-1.6	0.0	-0.2	0.0	0.2	0.1	-0.1	0.1	0.0	0.0	0.0	-0.2	3.7		0.0
	DR1	-0.1	0.3	0.0	0.4	0.0	0.7	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.6	-0.1	

Optimisation Analysis

One of the challenges of this study was the ability to be able to identify the optimal solution given a large solution space (5^{20} possible combinations). Software called OptQuest, which is part of a software suite called Crystal Ball was used in order to find the optimal solution in terms of maximising net profit. OptQuest incorporates meta-heuristics to guide its search algorithm towards the best solution and uses a form of adaptive memory to remember which solutions worked well before and re-combines them in to new, better solutions. Since this technique doesn't use the hill-climbing approach of ordinary solvers, it does not get trapped in local solutions, and it does not get thrown off course by noisy (uncertain) data. It results in an optimal or near-optimal solution by just analysing a fraction of all possible outcomes.

In order to estimate the optimum net profit, the modelled number of covers from the stage 1 analysis was combined with the item choices made in stage 2, and further to this TGI Friday's provided all fixed and variable costs. The raw preference shares were calibrated based on the real world volumes for each of the items in the choice model and re-weighted to reflect the actual number of starters, main courses and desserts purchased in relation to the total number of covers. This was necessary as the proportion of respondents choosing a starter or dessert in the survey was much higher than in real life. Due to the multi-menu method it was pragmatic to use a parsimonious approach to calibrating the data so an aggregate external effect was used in order to align the choice data to real world volumes. Individual level respondent weighting was also

incorporated that took into account frequency of eating out (Stage 1) and frequency of eating at TGI Friday's (Stage 2).

The client was further able to provide information regarding how previous changes in menu prices had affected covers so a scale factor multiplier was applied to the stage 1 data to adjust for the sensitivity of the logit response curve relative to actual price sensitivity.

The client launched the new menu in one if its restaurants in January 2010 and figure 6 shows comparisons against the same restaurant the previous year and a control group of restaurants for a number of key metrics.

The figures in the chart are index scores and show the change in the 2010 metrics compared to 2009. Results are shown after 3 and 6 months and are extremely encouraging. In the first 3 months the number of covers in the test store increased by 18% vs. 2009 (6% in control stores), Sales were up 14% (6% in control stores) and net profit was up 31% (12% in control stores), which represents a 17% uplift in net profit verses the control stores. Spend per head has decreased as the optimal menu decreased prices in 12 of the 20 items with only 6 items increasing in price so this was not unexpected. The figures after 6 months are still very encouraging and the relative decrease in uplift may be attributed to external factors such as the Football World Cup. It should be noted that net profit was adjusted for uncontrollable costs (Utility rebates, Repair work etc).

Figure 6

	Test Store Jan-Mar '10	Control Stores Jan-Mar '10	Uplift	Test Store Jan-Jun '10	Control Stores Jan-Jun '10	Uplift
Average # of covers	118	106	11%	114	104	10%
Average total week sales	114	106	8%	111	104	7%
Average total week profit	**131**	**112**	**17%**	**124**	**108**	**15%**
Average spend per head (core food)	96	99	-3%	97	100	-3%
Average Customer sat. score	125	121	3%	117	120	-3%

Index score vs. 2009 results (100) - Profit adjusted for uncontrollable costs

FURTHER ADJUSTMENTS TO THE MODEL

One of the issues that has been raised with regards to Serial cross-effect models is the potentially large number of cross-effect terms that need to be estimated. In this study there were 80 cross-effect terms across the 5 models created and this may lead to over-fitting, illogical cross-effects and degrade predictions. A solution to this is to only include significant cross-effect terms in the model.

Analysis of the models in this study were conducted using Sawtooth Software's CBC/HB software and one of the outputs from this software is the _ALPHA.csv file. This file contains all the random draws of the mean of the population distribution for each term. After removing the initial 10,000 burn-in iterations a term was classified as significant if 95% or more of the alpha draws were in one direction i.e. greater than zero or less than zero.

Using this test for significance, 35 of the 80 cross-effect terms were classified as significant and the models were re-run using this reduced set of cross-effect terms. Figure 7 shows a summary of the cross-effects in this pruned model.

Figure 7

		Model				
		Starter	Bundle	Main	Dessert	Drink
Cross –effect (Price)	Starter 1	N/A	✓	✓		✓
	Starter 2	N/A			✓	✓
	Starter 3	N/A	✓	✓		
	Starter 4	N/A		✓		✓
	Starter 5	N/A				✓
	Value meal 1	✓	N/A		✓	
	Value meal 2		N/A			
	Main 1		✓	N/A	✓	✓
	Main 2		✓	N/A		✓
	Main 3		✓	N/A	✓	
	Main 4		✓	N/A		✓
	Main 5		✓	N/A		
	Main 6			N/A		✓
	Main 7		✓	N/A		✓
	Main 8			N/A		✓
	Main 9	✓	✓	N/A	✓	✓
	Main 10	✓		N/A	✓	
	Dessert 1			✓	N/A	
	Dessert 2				N/A	✓
	Drink		✓			N/A

All the model diagnostics were re-calculated and compared against the model containing the full set of cross-effect terms – shown in figure 8.

Figure 8

	Original full cross-effect model		Pruned cross-effect model		**Improvement**	
	Holdout 1	Holdout 2	Holdout 1	Holdout 2	**Holdout 1**	**Holdout 2**
Mean Absolute Error (MAE)	0.48%	0.82%	0.46%	0.77%	**4%**	**6%**
Percentage Absolute Error (PAE)	5.40%	10.87%	5.22%	10.16%	**3%**	**7%**
R^2	0.992	0.956	0.992	0.966	**0%**	**1%**

In most cases there were moderate increases across all diagnostic measures when using the pruned cross-effect model. While these are unlikely to be significant, directionally it shows that the pruned cross-effect model may be a better and cleaner solution.

The optimisation analysis was re-run on the pruned model and in 5 of the 20 items, the optimal prices changed. Of these five changes, three were to increase price and two to reduce price. Of significance, in the Desserts model, only 6 of the 18 cross-effect terms were significant and this resulted in an increase in price of desserts of £1.50 from the original cross-effect model solution.

Acknowledgement

The author would like to acknowledge the contribution of Bryan Orme with regards to the guidance provided throughout the study.

References:

Bakken, David G. and Megan Kaiser Bond (2004), "Estimating Preferences for Product Bundles vs. a la carte Choices," Sawtooth Software Conference Proceedings, pp 123-134.

Cohen, Steven H. and John C. Liechty (2007), "Have it Your Way: Menu-based conjoint analysis helps marketers understand mass customization," Marketing Research Magazine, Fall 2007, pp 28-34.

Eagle, Thomas C (2009), "Turbo CBC: Volumetric Modeling," Turbo CBC Workshop, Anaheim, California, October, 2009.

Gilbride, Timothy and Greg M. Allenby (2004), " A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," Marketing Science.

Hauser, John R., Ely Dahan, Michael Yee, and James Orlin (2006) " 'Must Have' Aspects vs. Tradeoff Aspects in Models of Customer Decisions," Sawtooth Software Conference Proceedings.

Johnson, Richard M., Bryan Orme, and Jon Pinnell (2006), "Simulating Market Preferences with 'Build Your Own' Data" Sawtooth Software Conference Proceedings.

Johnson, Richard M., and Bryan Orme (2007), "A New Approach to Adaptive CBC", Sawtooth Software Research Paper Series.

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, May 2001, pp 183-196.

Orme, Bryan K (2000), "Hierarchical Bayes: Why All the Attention", Sawtooth Software Research Paper Series.

Orme, Bryan K (2007), "Advanced CBC Designs", Sawtooth Software Conference Tutorial

Orme, Bryan K (2010), "Task Order Effects in Menu-Based Choice", Sawtooth Software Research Paper Series

Orme, Bryan K (2010), "Menu-Based Choice Modeling Using Traditional Tools", Sawtooth Software Conference

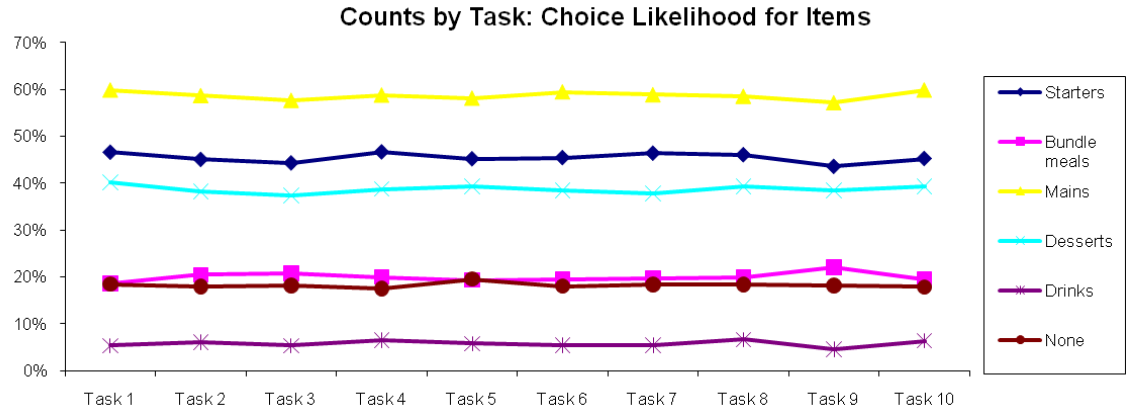
Sawtooth Software Inc, (2010), "Report on Conjoint Analysis Usage among Sawtooth Software Customers", Sawtooth software data.

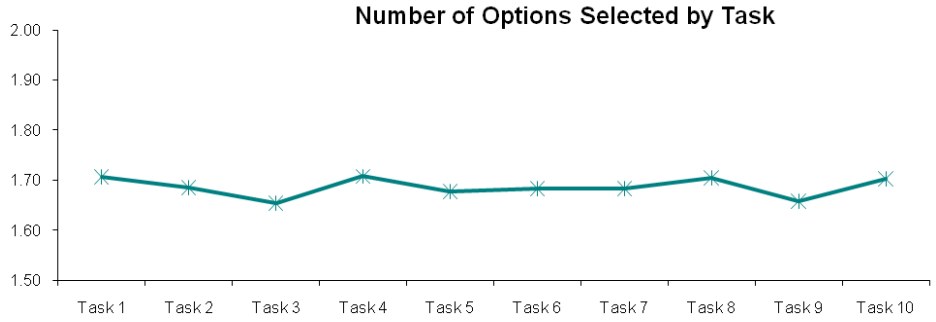
Sawtooth Software Inc, (2009) "The CBC/HB System for Hierarchical Bayes Estimation Version 5.0 Technical Paper", Sawtooth Software Technical Series.

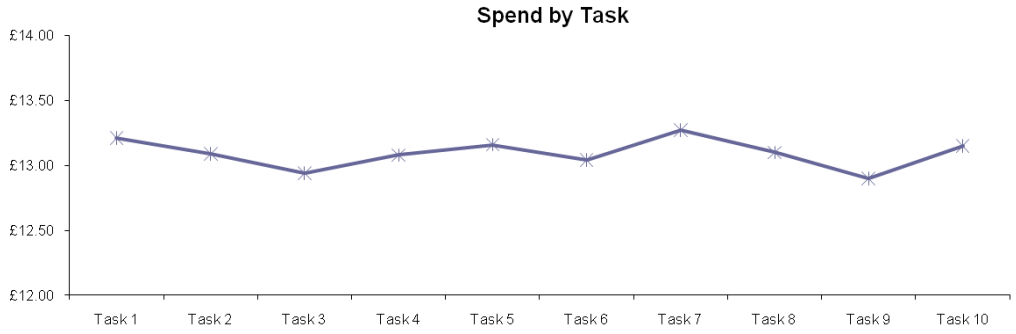
Wind, Jerry and Vijay Mahajan (1997), "Issues and Opportunities in New Product Development: An introduction to the Special issue," Journal of Marketing Research, 34 (1), 1-12.

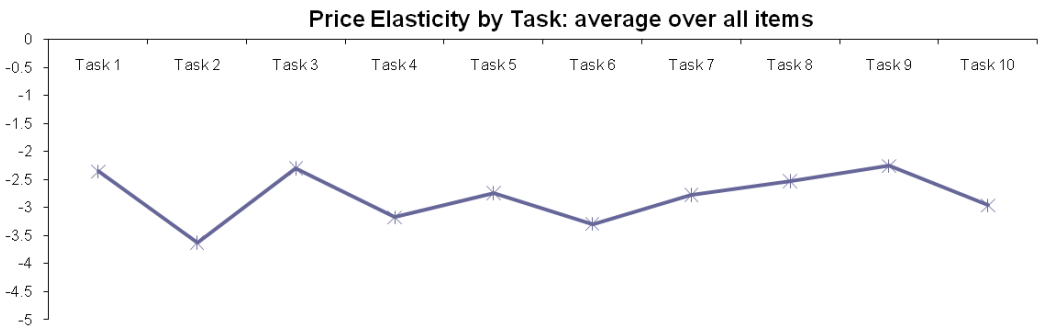
APPENDIX

Prior to the modelling, analysis was conducted looking at raw counts of key metrics by task.



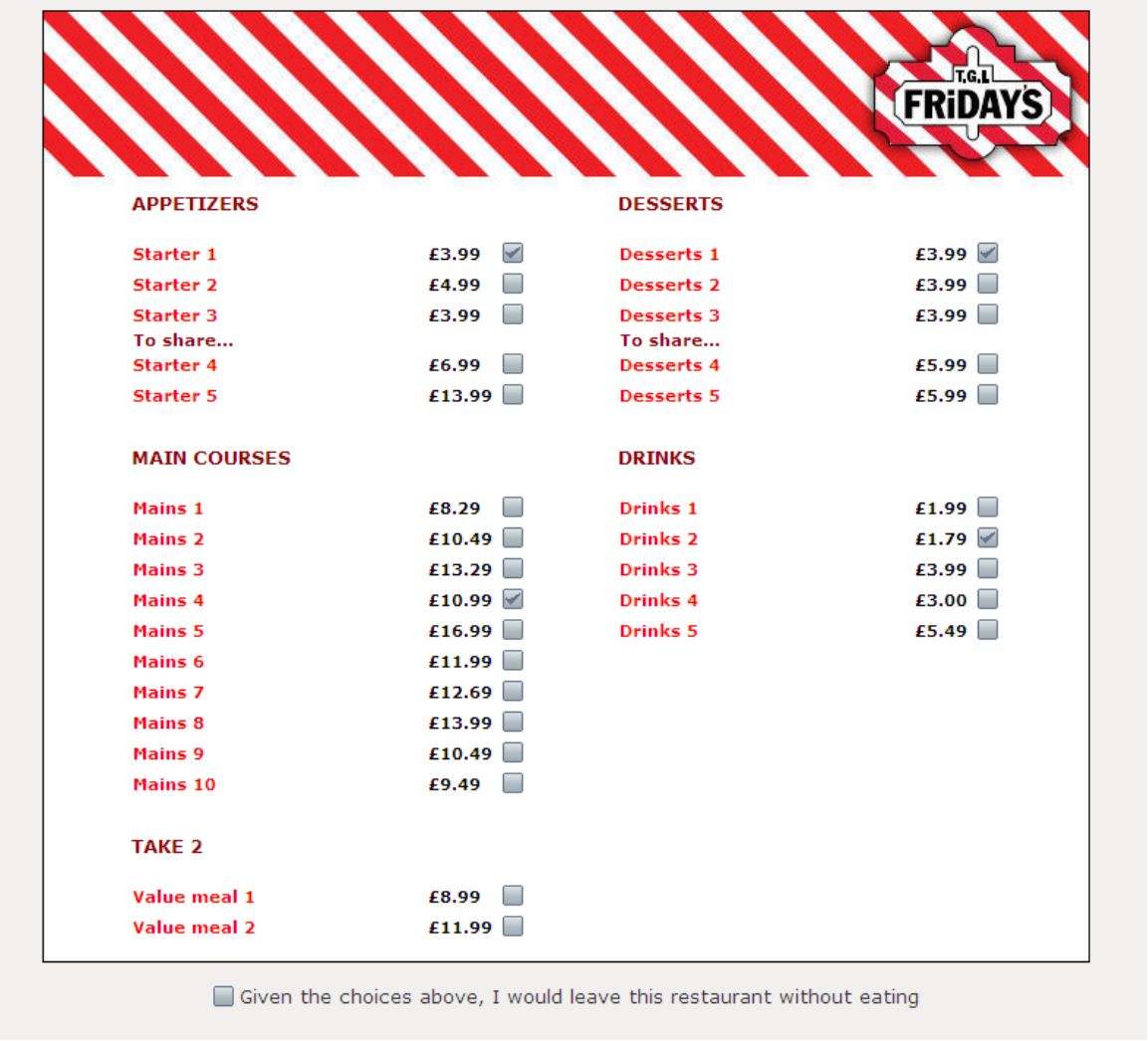






Across each of these metric there appears to be no pattern across tasks. One hypothesis for this is that respondents have already completed a CBC exercise so any learning effects have already been significantly reduced or eliminated. Orme (2010) observed significant increase in price elasticity as the number of tasks increased but in his case there was no prior exercise so learning effects may have been present.

An example of the choice task that respondents answered in the Menu-Based Choice stage is below.



AN EMPIRICAL TEST OF BUNDLING TECHNIQUES FOR CHOICE MODELING

JACK HORNE,
SILVO LENART,
BOB RAYNER,
AND PAUL DONAGHER
MARKET STRATEGIES INTERNATIONAL

INTRODUCTION

In some markets consumers can put together their own package of products or services for purchase. Examples abound and include services such as home cable/internet/telephony packages, as well as products as diverse as online book purchases, travel packages, PCs, and automobiles.

From a business perspective, however, companies are often unclear about how they will go to market. For instance, they could offer one or more fixed bundles of available features; or they might offer a "base" package to which consumers can add optional features at an incremental price; or consumers may be given complete freedom to build their own package from scratch where each "component" is priced separately.

Choice modeling seeks to mimic as much as possible the choice environments consumers are placed in for a given product category. At least two design/analytical techniques are available to applied market researchers with which bundling market scenarios can be addressed. These are:

- Fixed bundle choice-based conjoint (CBC) – In this approach, respondents choose one package from sets of differently configured product packages. The presence or absence of features is controlled using availability levels. A "no-buy" option is often included as part of the exercise. Prices may be shown explicitly for each feature, or as only a single price for each package.
- Build your own (BYO) – here, respondents are shown sets of available features each of which is offered at an explicitly shown price. Respondents build a package by choosing only those features that they want and will purchase together.

Other researchers have already drawn some comparisons between these methods. Bakken and Baker (2001) contrasted BYO and CBC techniques using an overall price feature in the discrete choice method that was completely uncorrelated with any of the other features. In subsequent work, Rice and Bakken (2006) used "latent" incremental feature prices that summed to a total price in comparing CBC to BYO, where total price was not used as an analysis variable. Finally, Johnson, et al. (2006) employed a similar CBC method where incremental feature prices were explicitly shown to respondents and summed to a total price.

In this paper we look at bundling from the perspective of the firm. With this focus, our paper builds on the extensive work of David Bakken and others (Bakken and Bond, 2004; Bakken and Bremer, 2003; Rice and Bakken, 2006).

In their analytical approach, Bakken and Bond conceive of consumer decision-making as a two-step process: "The buyer first chooses any one of the bundled offers or rejects all of the bundles. If all of the bundles are rejected, the buyer then chooses one or more of the components from an a la carte menu."

We have extended this work by employing a discrete product choice approach in which respondents are presented with choice tasks where alternatives consist of bundles defined by different combinations of the features. In any particular bundle alternative, a feature may or may not be shown; and if it is included in the bundle, it is shown at prices that vary. Essentially the approach uses availability effects to estimate the respondent's intrinsic utility for each feature, conditional on the set of price points at which the feature is offered.

We compare and contrast this methodology with the binomial-choice approach developed by Liechty, Ramaswamy and Cohen (2001) for BYO data. We believe their approach is especially well suited for analyzing menu-based data.

By comparing and contrasting empirical results from these two approaches, our study aims to inform market researchers on best practices for bundling studies so that they, in turn, can advise business decision makers on optimal strategies for moving bundled products to market.

Market Research and Bundling

The question of whether and how to bundle potentially complementary products or services is integral to market strategy for many firms and across many sectors (e.g., information technology, travel, entertainment/communications). Sometimes, this takes the form of standalone multifunction products such as printers/copiers/scanners. In other cases, the bundles may consist of separate products/services which can be sold as a package (e.g., flight/hotel/rental car).

For market researchers, bundling poses some very interesting challenges in the estimation of optimal bundle configurations and attendant pricing. One critical challenge is in deciding how to present choices to research participants to best inform the firm on their go-to-market strategies – should we present respondents with sets of fixed bundles and ask them which bundle they would buy or should we allow respondents to put together their own packages from a list of products/services?

We began this inquiry by posing (and expecting to support) the very simple null hypothesis that there will be no difference between a fixed bundling approach and a build-your-own (BYO) method. Assuming that respondents are rational, we should expect that the two methods will produce comparable results for bundle preference share and revenue because, quite simply, respondents want what they want and the task of uncovering what respondents really want is the singular aim of both approaches.

From the standpoint of bundling theory however (see Bakken and Bond, 2004), we could just as easily argue that respondents would be more sensitive to their reservation prices – the price each respondent is willing to pay – for individual items when those items were presented as individual choices from a BYO menu than when they were presented in fixed bundles. If this were the case, since BYO places a greater focus on individual items while fixed bundling privileges the package as a whole, we would expect take rates for items to be higher in the fixed bundling approach, and consequently, greater revenues for the firm.

Our main question, therefore, was as follows: When applied to the same market research question, are the fixed bundle (CBC) and BYO approaches interchangeable or are they inherently different? This question has both methodological and substantive implications:

1. Methodological – All design criteria being equal, do respondents react and respond to CBC and BYO tasks differently? If so, how should modeled results from the two approaches be interpreted?
2. Substantive – From the perspective of the firm, will one approach be better in providing insight into how consumers respond to bundling and inform optimal go-to-market strategies?

In order to test these questions, we focused our research on a marketplace that has included very robust practitioners of bundling, namely home-based entertainment and communication services. Here, providers offer packages of services that span from landline telephones and internet access, to mobile phone and wireless internet provision, to cable and satellite television. The fact that these disparate services are now pervasively offered across the country means that it is a marketplace where most consumers have knowledge of and experience with bundling. Consequently, it provides us with an excellent case study for our research approach questions.

Bundling Approach and Design Considerations

In the course of designing our research, it became apparent that we must be mindful of the specific market idiosyncrasies that may impact bundling dynamics. So, for instance, our choice of the home entertainment/communications market carries with it some very important considerations.

To begin with, this marketplace is characterized by robust competition among firms that have different origins and legacy strengths. Among these are traditional landline companies like AT&T; subscription television entertainment providers based on cable or satellite technologies such as Comcast and DirecTV; and relatively newer mobile communications providers such as T-Mobile and Sprint. All of these firms either already offer, or have the potential to offer, all of the services mentioned. Further, consumers have the ability not only to stay (and bundle) with one provider, they can also pick and choose providers across individual or mini-bundles of services. A consumer might purchase cable television from one provider, landline and home internet from a different provider and mobile services from yet another provider.

Given these characteristics, a market in which both fixed bundles within providers, as well as, build-your-own packages across providers is not only possible but fairly common. Assessing how well each of the two approaches perform in estimating key market measures (preference share, revenue) becomes even more palpable for the research community.

On the question of properly modeling a competitive marketplace, a second consideration became apparent: How well does a "traditional" BYO exercise model competition? The default, or what we term "traditional," implementation of a BYO is as follows: Respondents are shown one set of features at a time and they pick the ones they want from that set. Typically, the solution for modeling a competitive marketplace is to vary brand across successive BYO tasks or to have respondents choose brand among the other menu selections. The experimental design, it is assumed, will result in brand specific responses that allow for modeling the competitive landscape.

Clearly, the straightforward way to model the competitive communication/entertainment market via BYO is to present respondents with all relevant brands in each task with each brand shown with its own available/not available services, and allow respondents to pick the services they want from the various brands. To be sure, this "market BYO" variation seems a bit daunting to implement. If we have a large number of competitors, respondents are subjected to a very complicated, time- and attention-consuming set of exercises. All the same, since the "market BYO" variation is more directly comparable to the fixed bundle choice set, we decided to include it as a variation on the more "traditional" BYO approach in our test of bundling methodologies.

A Case Study: Home-based entertainment and communication services

Sample and design

A total of 1,502 US consumers recruited from the eRewards panel participated in an online survey. Key screening criteria included age and subscription (or intent to subscribe) to one or more home-based entertainment and communication services (including landline phone, mobile phone, home-based high-speed internet, mobile high-speed internet, cable TV and satellite TV). Respondents were randomly assigned to one of three experimental groups:

1. Fixed Bundles where choice tasks consisted of sets of pre-configured bundles of services offered by different brands (n=501);

2. Single Brand BYO where choice tasks consisted of pre-configured sets of individual services offered by a single brand at a time (n=499);

3. Market BYO where choice tasks consisted of pre-configured sets of individual services offered by different brands[1] (n=502)

All services had "not available" levels and three specified price levels (low, mid, high). Overlapping designs were used for the services, such that all of the non-TV services were available 3/5 of the time and TV was available 6/7 of the time. Since cable and satellite TV could not both be available from the same brand in any given choice task, they were each available 3/7 of the time. Six brands, representative of those that are currently offering such services in the US were included as a non-overlapping feature.

Finally, a discount feature with four levels – none, low, mid, high – was included for each possible configuration. The applied discount for each brand was the product of the discount level in the design and the number of services available after the first (Fixed Bundles) or the number of services selected from a single brand after the first (Market BYO and Single Brand BYO). In the two BYO conditions, the applied discount, along with the total subscription cost, automatically updated as selections were made.

In all three conditions, respondents had the option to select "none of these bundles" or "none of the items available." Figure 1 shows an example of a typical Market BYO task.

Identical designs were used for the Fixed Bundles and Market BYO conditions. A unique design was generated for the Single Brand BYO condition since only one configuration from a

[1] While all services could theoretically be available from all brands in the Market BYO condition, respondents' choices were restricted so that a given service could only be selected from a single brand.

single brand was shown in a given choice task. In all designs, there was an average of 3.26 services available per brand. All respondents in all conditions evaluated 12 choice scenarios.

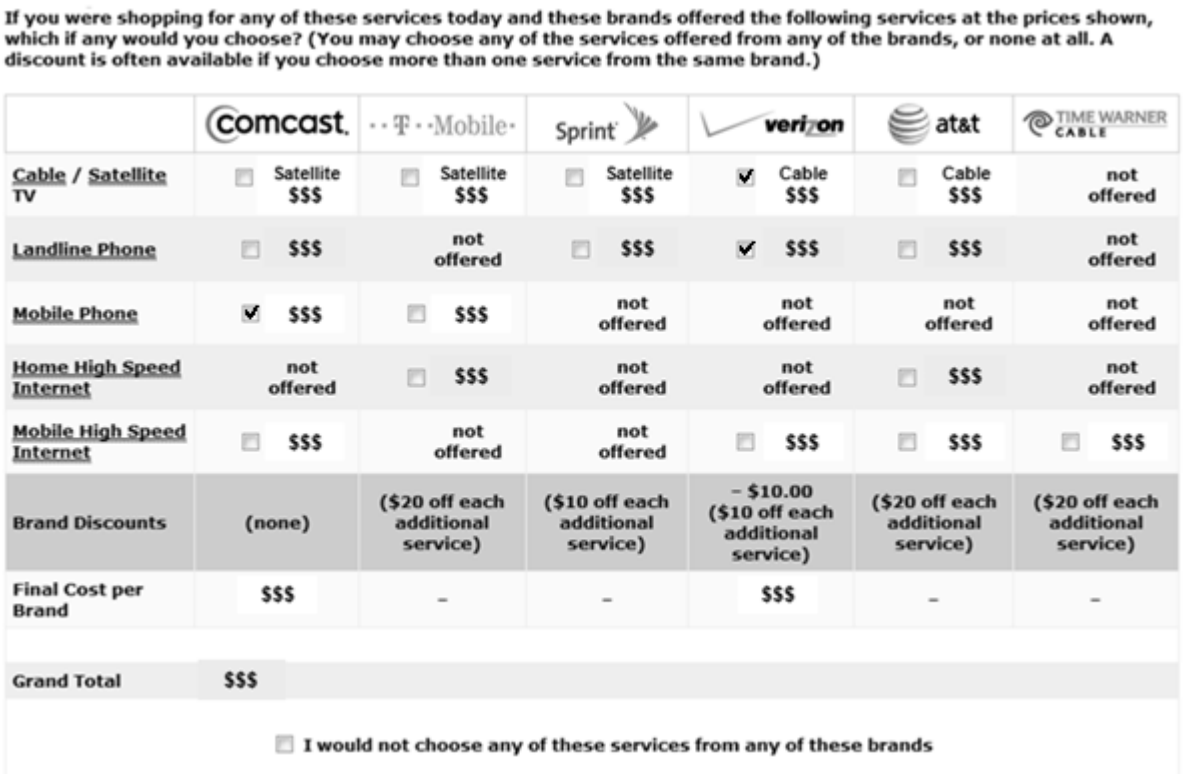


Figure 1: Example of a Market BYO choice task

Prior to the first choice task, respondents saw an education screen identifying what was included in each of the services (e.g., "Mobile Phone: Individual plan (family plans would be available), free Mid-Range phone, 1000 minutes of talk time with unlimited nights and weekends. Unlimited text. Internet and additional calling minutes and features available at additional cost."). The mid prices used for each service in the design were the approximate current market prices of the services described on this education screen. Low prices were 30% less than these values and high prices were 50% more. All prices were generic to brand.

The screener was minimal and required an average of 2 min to complete. Following the choice scenarios were five self-evaluation questions that used the same language as those found in Orme (2010), and several demographics.

Description of the models

All choice data were modeled at the individual level using CBC/HB. Part-worths were estimated for a buy/no buy intercept, six brand levels, four levels for each of the non-TV services, four levels for discount, and for TV: a three-level variable indicating availability – cable, satellite or neither available – and a three-level variable indicating price – zero coded if neither cable nor satellite was available for a given brand. Individual level data were converted to "first choice" in all models after finding the utility associated with each possible selection.

Additional details for each model are as follows:

Fixed Bundles: A single model was calculated using all of the above parameters. Utilities were calculated for the selection of each brand (i.e., bundle offered) and for a "none" option.

Market BYO: Six models, one for each service, were calculated. Cross effects from other services (i.e., services not being modeled) were converted to binary, available/not available, variables prior to modeling. Including the full cross-effects (i.e., availability and price) resulted in substantial noise among the estimated parameters, so the full cross-effects models were discarded as not useful early on. Full price effects for the service being modeled, along with full discount effects, were included in each model. For each service, utilities were calculated for the selection of each brand and for a "none/not selected" option if a given service was not selected at all on a given choice card.

Single Brand BYO: Same modeling approach as in Market BYO, except only two utilities were calculated for each model: one for the selection of the service being modeled and one for the non-selection of the same service.

Descriptive results

Market BYO and Fixed Bundles tasks consistently required more time to complete than the Single Brand BYO tasks (Figure 2). In all three conditions, the first choice task required the longest time to complete with Market BYO tasks averaging 77 sec, Fixed Bundles averaging 60 sec, and Single Brand BYO averaging 40 sec. After about the fourth task, these averages settled to just under 30 sec per task for Market BYO and Fixed Bundles and about 15 sec for Single Brand BYO. The trend of Market BYO > Fixed Bundles > Single Brand BYO was present at each task order.

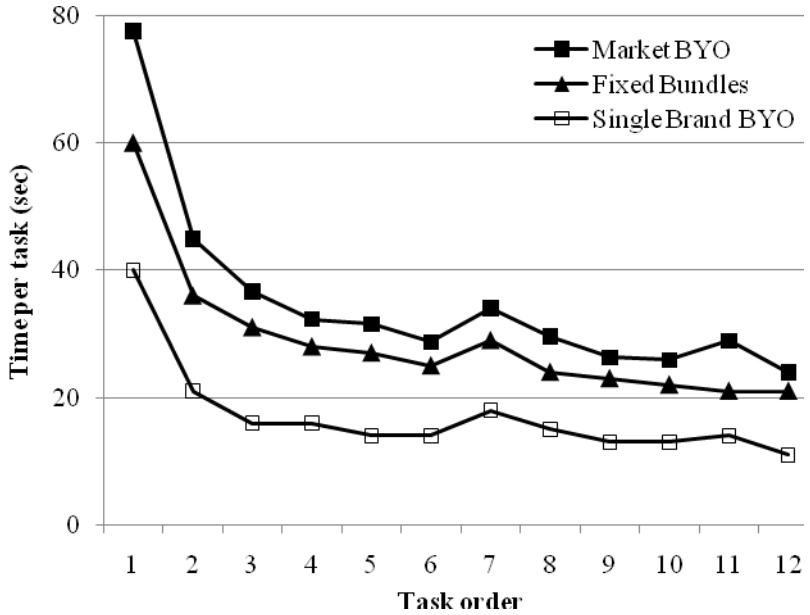


Figure 2: Time to complete choice tasks by method and task order

More than three times as much revenue was generated in individual Market BYO and Fixed Bundles tasks ($81 and $76 per task, respectively) than in Single Brand BYO ($21 per task). These revenue calculations factored in zeroes for those tasks where "nothing/none of these" was

selected, and this accounts for some of the differences between the methods. In the Single Brand BYO condition, respondents did not select a single service on nearly 65% of all choice tasks. We attribute this to a favorite brand effect. Since only a single brand was present on each choice card, a given brand was only available to respondents on 2/12 (17%) of all choice cards. If respondents had a tendency to only choose items from a preferred brand, this high rate of choosing "none" makes sense. However, even on choice tasks where respondents did choose one or more services, they tended to choose fewer than two in the Single Brand BYO (avg. 1.6 services per task). This was in contrast to Market BYO and Fixed Bundles where respondents selected an average of 3.0 and 3.1 services per task, respectively, when they chose anything at all. Thus, the trend of more revenue being generated in the latter two methods persisted, even when zeroes from cards where nothing was selected were not factored into the calculations.

Respondents chose "none of these" on 55% of Fixed Bundles tasks and 25% of Market BYO tasks. This difference is likely due to the fact that respondents had to commit to everything in a bundle in Fixed Bundles tasks, while they could take just one item from a given brand (all of which were always available) in Market BYO. Along the same lines, respondents were twice as likely to select five services in any single choice task (the maximum possible) in Market BYO (14% of tasks) than in Fixed Bundles (7% of tasks). Respondents selected five items in only 2 tasks out of nearly 6,000 total in the Single Brand BYO condition. The average number of services selected per task was flat across task order in all three conditions.

Market simulations

To test the results of the models on shares and revenue generation, we ran several identical market level simulations for each experimental condition. Within these simulations:

- All services were available from all brands (cable TV was offered from the two brands already offering cable TV in the marketplace, satellite TV as offered from the other four brands).
- Pricing and discounts were set as follows:
    - All services offered at low prices, high discount
    - All services offered at mid prices, low discount
    - All services offered at high prices, no discount

In addition to shares of preference and revenue across brands in each simulation, we measured total market reach and revenue for each condition. For the purposes of these analyses, these concepts were defined as follows:

Market Reach (Fixed Bundles): Proportion of respondents selecting any of the fixed bundles in a given simulation (i.e., proportion not selecting "none of these").

Market Reach (BYO conditions): Proportion of respondents selecting at least one service from any brand in a given simulation (i.e., proportion not selecting "none of these").

Market Revenue: Total revenue generated across all brands measured in dollars per 100 respondents per month.

Market Reach (Figure 3, left panel) was highest in the Market BYO and lowest in Fixed Bundles at all three price levels. The range across these values was not extreme at low

prices/high discount (Market BYO: 84%, Single Brand BYO: 79%, Fixed Bundles: 78%), but became much larger at high prices/no discount (Market BYO: 53%, Single Brand BYO: 28%, Fixed Bundles: 20%). At low prices, then, it seems that the favorite brand effect previously noted in the Single Brand BYO played much less of a role. Respondents were nearly as likely to take services from a less preferred brand as long as those services were offered at a low price. At high prices, the favorite brand effect was very much in play. Further, at high prices, respondents became much less likely to select a complete bundle. In contrast, Market Reach for Market BYO stayed relatively high, even at high prices and no discount. In this condition, just over half of respondents could still find one or more services from one or more brands that they were willing to select.

Market Revenue (Figure 3, right panel) was highest in all three conditions at mid prices/low discount. The same relationship persisted between the conditions across all three price levels (Market BYO > Fixed Bundles > Single Brand BYO), and again, the differences were not extreme at low prices/high discount (Market BYO: $6,899, Fixed Bundles: $5,975, Single Brand BYO: $5,787). However, at mid and high prices, these differences became much larger, especially the differences between revenue from the Single Brand BYO and the other two conditions. At mid prices/low discount, Single Brand BYO generated $6,000 in revenue (3.6% increase in revenue compared to low prices/high discount), while Market BYO and Fixed Bundles generated $9,712 (+40.8%) and $9,421 (+57.6%) in revenue, respectively.

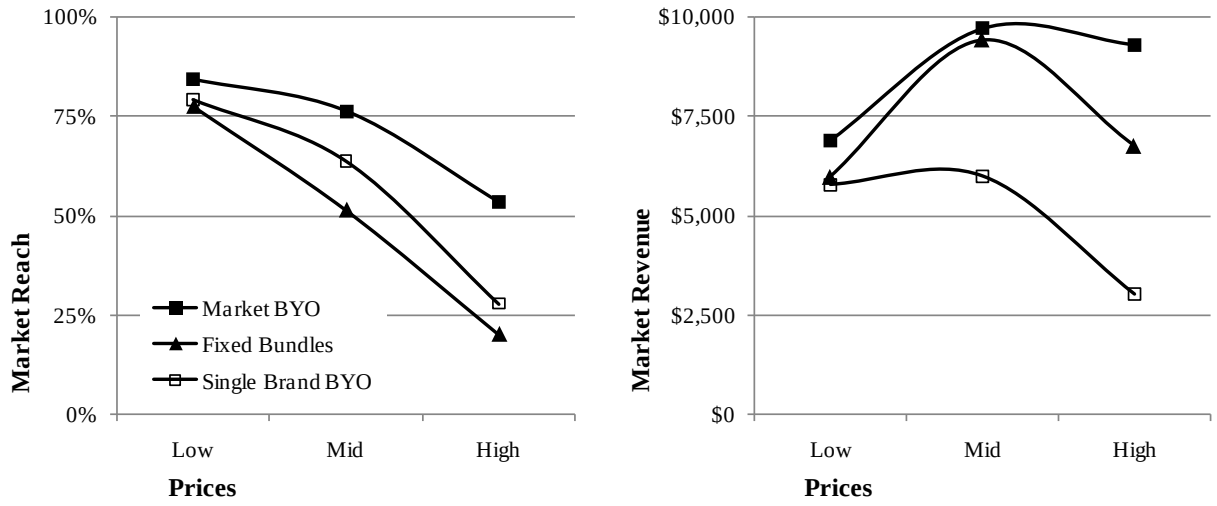


Figure 3: Market reach (left panel) and market revenue (right panel) for the different methods at various price points for the individual services

The revenue trends persisted when looking at individual brands. That is when the prices of offerings from a single brand were manipulated and those from all other brands were held at mid prices and low discount, Market BYO nearly always generated the most revenue for the individual brand that was manipulated and Single Brand BYO nearly always generated the least. The shapes of these revenue curves however were not the inverted "U" shapes seen in the market as a whole. Rather, the most revenue was always generated at the lowest prices since that is the

level where the brand under consideration is out-competing all of the other brands in a given simulation.

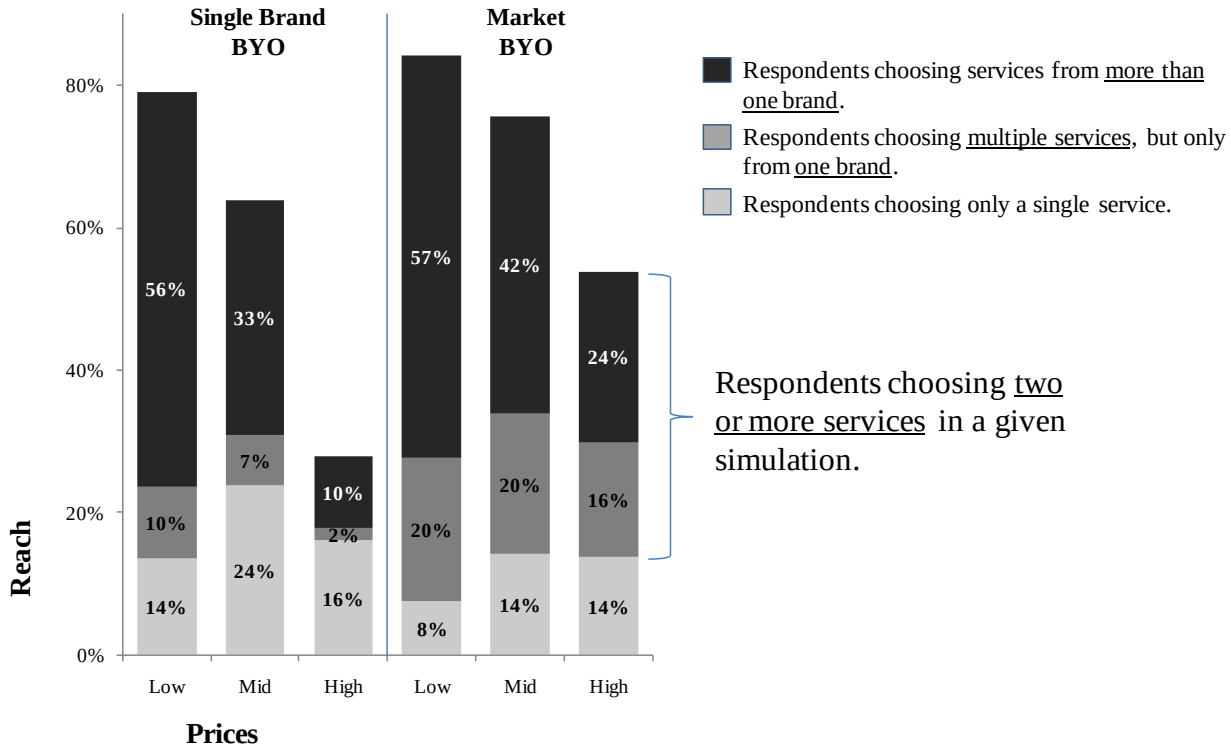


Figure 4: Number of services selected in market simulations of BYO conditions

With respect to the number of services selected, in the Single Brand BYO, respondents not only selected "none of these" more often as prices increased, but they selected fewer services overall in those tasks where they selected something. Figure 4 shows the relationship between the two BYO conditions in terms of the number of services selected in each simulation and the number selecting services from multiple brands. First, in the Single Brand BYO, it was highly unlikely to find respondents selecting more than one service from just one brand. Only 10% of respondents did this at low prices/high discount, and the number fell to 2% at high prices/no discount. More than twice as many respondents did this in the Market BYO condition at each price level. Further, at high prices/no discount, the ratio of respondents selecting two or more services to those taking just one service was nearly 3:1 (40%-to-14%) in the Market BYO condition, compared to 0.75:1 (12%-to-16%) in the Single Brand BYO. If respondents are apt to take multiple services from multiple brands in both BYO conditions, they are more likely to do so when presented with a full array of services available across brands than they are when presented with just one brand at a time. This appears to be the case especially as price levels increase.

We can see this occurring at the individual service level in Figure 5. At mid prices/low discount, the selection rate for individual services was nearly always lowest in the Single Brand BYO, while the same selection rates in Market BYO and Fixed Bundles were more similar to one another. The differences between the conditions were especially large for the mobile broadband internet service. This is the latest service out of all of the ones tested to emerge in the

marketplace, and likely has the lowest current penetration. The low selection rate for this service in Single Brand BYO (8%) is perhaps indicative of its low current penetration. But when offered as part of a fixed bundle, more than six times as many respondents selected a bundle containing the service (51%). This seems to be a strong argument for offering new-to-market products and services in packages with other high-penetration and well-known products and services. Doing so will increase awareness, market penetration and ultimately revenue from the new-to-market offerings.

The relationships between the conditions were reversed somewhat for TV. Taking cable and satellite TV in aggregate, the highest selection rates were in the two BYO conditions (both at 59%) and lowest in Fixed Bundles (51%). We suspect that the reason for this is because consumers may tend to conceptualize TV as a "traditional" service distinct from telephony and internet access, and so be more likely to subscribe to it on its own.

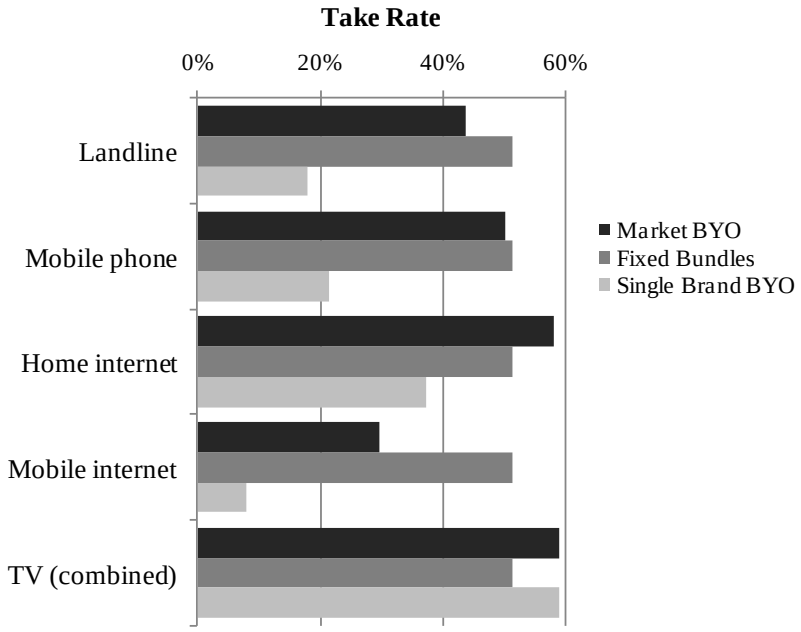


Figure 5: Take rate of individual services from market simulations. All services offered at mid prices

Price sensitivities were higher for all services except mobile broadband internet in the two BYO conditions than they were in the Fixed Bundles condition (Figure 6). In the case of mobile broadband internet, this again could be due to the newness of this service to the marketplace. Relatively few respondents selected mobile broadband internet even at low prices, so at high prices, the selection rate could not fall much further.

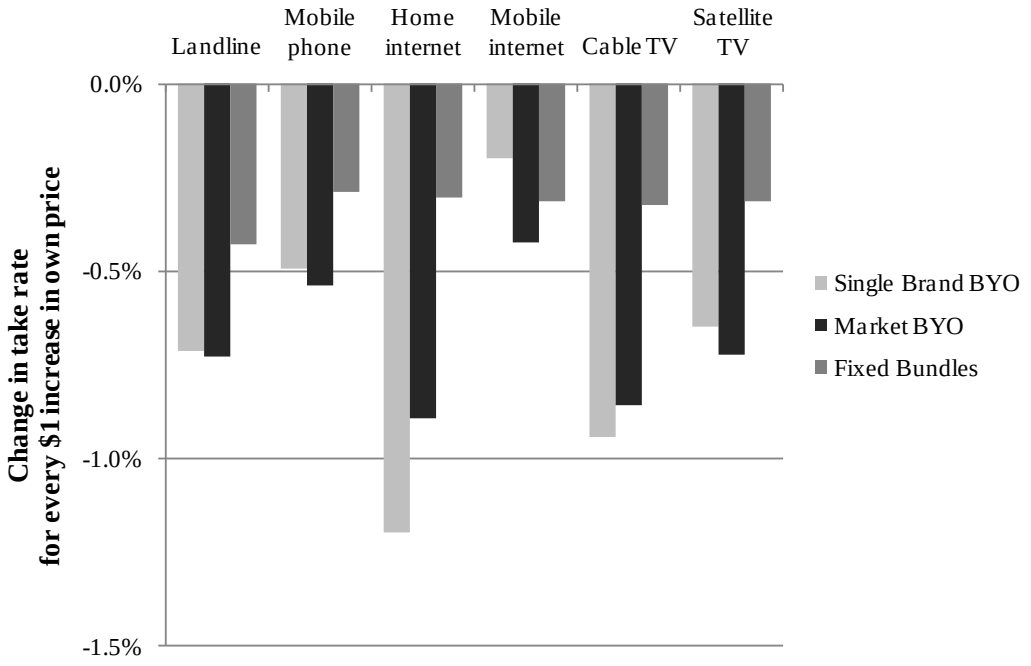


Figure 6: Price sensitivity by service and method

The differences between price sensitivities in the BYO conditions and the Fixed Bundles condition among the other services can be explained by the different focus the respondent likely has when participating in the different types of exercises. In BYO conditions, where the respondent has to actively select each service or product, he or she pays attention to the individual service or product prices. In contrast, when participating in a Fixed Bundles task, where the respondent is not actively selecting each service or product, but an entire bundle at once, it appears that he or she pays more attention to the total cost of the bundle, even if the prices of the individual services or products are still shown.

Self evaluation of the exercises

Respondents rated the three experimental conditions to be fairly similar across five metrics asked at the end of the survey (Table 1). Still, there were a few interesting differences among the methods. Market BYO and Fixed Bundles were more likely to be rated "monotonous and boring" than Single Brand BYO, and Single Brand BYO generated more "interest in taking another survey just like this in the future." Further, Single Brand BYO was rated as net "better" than other internet surveys, while Market BYO was rated as net "worse." These results were likely due to ease of processing and thus the speed at which the respondents were able to get through the Single Brand BYO tasks compared to the other two methods. Respondents may prefer exercises they can get through quickly and without expending a lot of mental energy. This, however, doesn't necessarily mean that they provide better results in such exercises. In fact, respondents agreed that "the way these services were presented made me want to slow down and make careful choices," more often in the Fixed Bundles condition than in either of the two BYO ones. As to our concern that Market BYO tasks may be comparatively tedious, this is borne out by the "overall experience" results – respondents in this condition tended to rate their experience as worse than other types of surveys.

Table 1: Summary from self-evaluation attribute ratings. All attributes were rated on 5-point scales. See Orme (2010) for exact wording of questions used. For each question, dark shading represents top method in each column; and light grey shading indicates bottom method in each column.

	Attribute	Top 2 Box	Bottom 2 Box	NET	Mean
Single Brand BYO	Monotonous & boring	31%	30%	+1%	3.01
	Interest in taking similar surveys	55%	14%	+41%	3.50
	Easy to give realistic answers	54%	22%	+32%	3.38
	Slow down and make careful choices	45%	17%	+28%	3.31
	Better than other online surveys	34%	20%	+13%	3.17
Market BYO	Monotonous & boring	55%	17%	+37%	3.47
	Interest in taking similar surveys	49%	24%	+26%	3.29
	Easy to give realistic answers	48%	26%	+22%	3.24
	Slow down and make careful choices	50%	22%	+28%	3.32
	Better than other online surveys	26%	35%	-9%	2.87
Fixed Bundles	Monotonous & boring	47%	17%	+30%	3.34
	Interest in taking similar surveys	48%	20%	+29%	3.32
	Easy to give realistic answers	54%	22%	+33%	3.37
	Slow down and make careful choices	62%	18%	+45%	3.53
	Better than other online surveys	30%	24%	+6%	3.06

Finally, survey abandon rates are another form of self-evaluation. If more respondents abandon a survey prior to completing, then we have to burn through more of them as a whole to get to our desired number of completes. When targeting consumers, this is not a major issue, but when targeting harder-to-reach groups, burning through more respondents can substantially raise our costs of doing research. There were some large differences in abandon rates between these conditions. Nearly five times as many respondents abandoned sometime after completing the first choice task in the Market BYO condition (19%) than in the Single Brand BYO (4%). Fixed Bundles fell in the middle, with an 8% abandon rate after the first choice task. Again, these numbers are reflective of the amount of time respondents spent in each of the exercises, and point to using some caution at design-time towards including a lot of exercises that respondents might find tedious or boring.

DISCUSSION

The results of our experimentally-controlled investigation into bundling approaches clearly indicate that, in the specific market we examined, BYO and Fixed Bundle methodologies are not interchangeable. From the perspective of our key initial questions, a quick summary of our main findings is as follows.

Methodologically, respondents react and respond to CBC (Fixed Bundles) and the two different kinds of BYO tasks differently. Respondents process and complete Single Brand BYO tasks more quickly and abandon the survey less frequently. This does not mean, however, that the resultant data are of better quality from this method. There is some evidence here that respondents rushed through this type of task more than the others, and in that sense the data might be considered to be of lesser quality.

The BYO exercises as a group led to greater price sensitivities for individual items than the Fixed Bundles method. This suggests that respondents focused more on individual component pricing in BYO tasks while, in Fixed Bundles tasks, they focused more on the price of the bundle as a whole. This particular finding appears to confirm prior bundling theory that reservation prices on individual items should exert a stronger effect if those items are offered separately than if those same items are offered as parts of a bundle.

Last, from the methodological standpoint, the Fixed Bundle and Market BYO approaches both resulted in higher predicted selection rates for individual items than the Single Brand BYO method. Different dynamics appeared to be present in the first two methods. In the Fixed Bundle approach, the higher selection rates came from respondents being forced to take some items that they might not have purchased if offered separately, simply because those items were part of an otherwise attractive bundle. In Market BYO, on the other hand, high selection rates resulted from being able to choose individual items from a variety of brands, rather than being restricted to just one brand in the Single BYO condition.

From the perspective of the firm, the CBC and BYO approaches provide different insights into consumer orientation and responses. As above, Fixed Bundles mitigate individual component price sensitivities, and for the firm, this might provide a valuable market entry point for new products or services.

There is some evidence here to suggest that the Market BYO method maximizes both market reach and market revenues, which might also extend to individual brands. This may be due to some consumers only wanting a single item and thus not being satisfied with any bundle containing two or more items. This type of individual would be "reachable," thus providing revenue, in the Market BYO framework, but not in the Fixed Bundles one.

The Single Brand (or unbranded) BYO approach does not appear to be useful in the marketplace we tested. Clearly a favorite brand effect was in place with this method, and otherwise attractive bundles were rejected in choice tasks merely because of the brand they were associated with. This method may be more useful in the early investigative stages of bundling when only a single brand, or no brand at all is being tested. It could also prove to be of greater use in markets where consumers can only purchase from a single brand at a time such as in the case of quick-service restaurants or automobiles.

Finally, we can conclude from this work that in bundling research, market idiosyncrasies do matter. Fixed Bundle (CBC) approaches are most appropriate in markets where products and services cannot be bundled across brands. Market BYO appears to be more appropriate in markets where both within and across bundling is possible. In-and-Out Burger can offer a combo meal for the same price as the component parts of the meal because I can't buy an In-and-Out milkshake and a MacDonald's hamburger at the same time. BestBuy on the other hand would probably want to think twice before bundling a camera, a computer and an MP3 player from the same manufacturer at the same price of the cost of the individual components.

Conclusions

We found that the default (Single Brand) BYO approach is not a proper tool for modeling bundling dynamics in the kind of competitive marketplace where consumers can take some items from one brand and other items from different brands. However, for purposes of pre-launch investigation and very simple single brand research, the Single Brand BYO provides an elegant and efficient option.

The choice of which of the remaining two bundling research approaches to use ultimately depends on the particular idiosyncrasies of the relevant marketplace. If items would likely be sold individually from different brands, the Market BYO approach may be considered the most appropriate research method, while if the marketplace doesn't support buying individual items across brands, the Fixed Bundles approach may be more appropriate.

References

Bakken, D. G. and L. R. Bayer (2001), "Increasing the Value of Choice-based Conjoint with 'Build Your Own' Configuration Questions." 2001 Sawtooth Software Conference Proceedings, Victoria, B.C.

Bakken, D.G. and J. Bremer (2003), "Estimation of Utilities from Design Your Own Product Data," A/R/T Forum, June 2003, Monterey, California.

Bakken, D.G. and M. K. Bond (2004), "Estimating Preferences for Product Bundles Vs. A La Carte Choice," Sawtooth Software Conference Proceedings, San Diego, California.

Johnson, R., B. Orme and J. Pinnell (2006), "Simulating Market Preference with 'Build Your Own' Data." 2006 Sawtooth Software Conference, Delray Beach, Florida.

Leichty, J. V., V. Ramaswamy and S. 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, Vol. 38, Number 2.

Orme, B. (2010), "Task order effects in menu-based choice." Sawtooth Software Research Paper Series (http://www.sawtoothsoftware.com/download/techpap/mbcorder2010.pdf)

Rice, J and D. G. Bakken (2006), "Estimating Attribute Level Utilities from 'Design Your Own Product' Data." 2006 Sawtooth Software Conference, Delray Beach, Florida.

ANCHORING MAXIMUM DIFFERENCE SCALING AGAINST A THRESHOLD – DUAL RESPONSE AND DIRECT BINARY RESPONSES

KEVIN LATTERY
MARITZ RESEARCH

I. INTRODUCTION

Maximum Difference Scaling is a choice based tradeoff technique for understanding the relative value of several attributes. Respondents are asked to choose the "best" and "worst" attribute from a subset of the attributes. An example of a MaxDiff question is this:

Thinking of your ideal Mobile Phone Retail Store, which of these features is most important and which is least important to you?

	Most Important	Least Important
Friendly sales representative	◉	○
Store front is attractive	○	◉
Convenient store location	○	○
Information about rebates/discounts	○	○

Respondents see several screens like this, each time choosing their best and worst attribute. This kind of technique is useful because it does not depend upon how respondents use a scale. Instead it asks respondents to make choices. So it is in the family of tradeoff techniques and shares similarities with conjoint analysis. Like a conjoint, MaxDiff also has an experimental design and can be analyzed using the same techniques as conjoint. Indeed the most common form of analysis is Hierarchical Bayes (HB) such as Sawtooth's CBC/HB module (which was what we used in this paper).

What we understand from MaxDiff is the relative value of each of the attributes. To make this point clearer, consider the following thought experiment. Imagine two respondents, call them Brad and Angelina, who would rank order the attributes the same way. They would therefore answer each MaxDiff task the same way, and as a result we would derive the same utilities for both of them (within error). But as it turns out, Brad and Angelina are very different. For Angelina, all of the attributes are important, while for Brad none of them matter. So while the rank order is the same, all of Brad's utilities should be shifted lower, in fact much lower than Angelina's utilities.

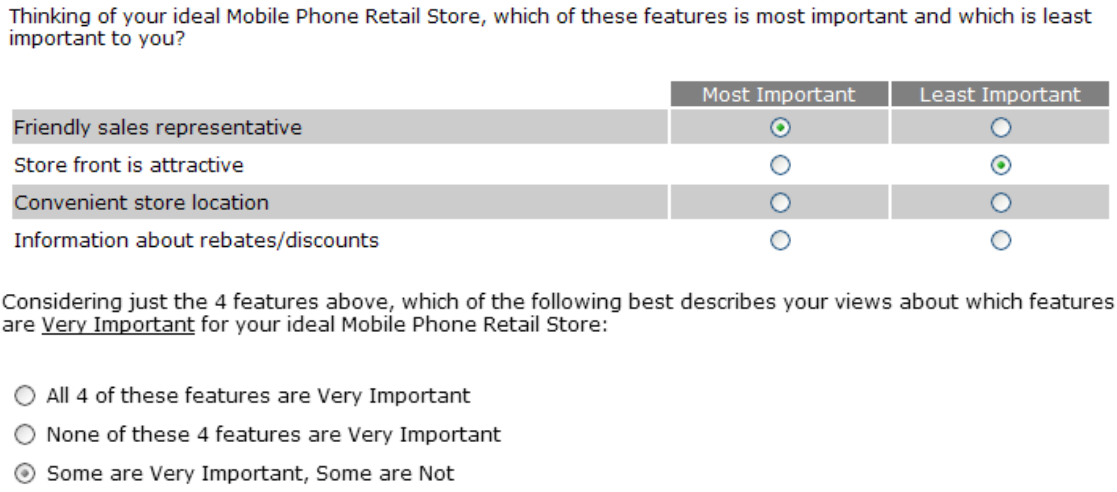
In some cases, relative utilities are fine, but sometimes researchers want the utilities to take into account some kind of absolute measure. That is, one may want Brad and Angelina's utilities to be different because Brad really thinks none of the attributes are important, while Angelina does.

One method to make MaxDiff utilities less relative is to anchor the utilities to a specific point. For instance, we might make a utility of 0 to be a reference point, where above 0 means important and below 0 means unimportant. This means all of Brad's utilities would be shifted below 0 while Angelina's would all be above 0.

This is known as anchoring the utilities to a threshold. In the example above 0 was a threshold to which the utilities were anchored. In the remainder of this paper, we will be discussing two methods for anchoring utilities to a threshold of 0: Indirect Dual Response and Direct Binary Responses.

II. Anchoring Technique One: Indirect Dual Response Method

This method was first suggested by Jordan Louviere. After each MaxDiff task, one asks a follow up question about whether all, none, or some of the attributes meet a threshold. An example of this follow up question is shown below:



In this case the follow up question asks whether the attributes are "Very Important", but any other phrase could be used. This will become the anchoring that corresponds with a utility of 0. So in this case attributes with a utility above 0 are "Very Important", while attributes with a negative utility are not "Very Important".

Implementing this method requires some clever coding. First, one no longer uses a reference level. For the best and worst pick, one uses full dummy coding. The example below will show how a specific task is coded. For this example, assume there are 8 attributes, and the respondent saw attributes 1, 3, 6, and 8.

For the best choice we have the following coding, which is the same as typical MaxDiff coding without a reference level:

a1	a2	a3	a4	a5	a6	a7	a8
1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1

For the worst choice we also have the typical MaxDiff coding but without a reference level:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	-1

The trickier part is how to code the follow up question.

If the respondent said “None are Very Important” then one added the following task:

a1	a2	a3	a4	a5	a6	a7	a8
1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0

We pretend the respondent saw 5 attributes, with the 5th fictional attribute winning. The idea here is that each of the attributes loses to the zero vector and therefore the utilities will be negative.

If the respondent said “All are Very Important” then we add the following task instead:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	-1
0	0	0	0	0	0	0	0

Again we pretend the respondent saw 5 attributes with the 5th fictional attribute winning. In this case, the negated attributes lose to the zero vector, meaning they are positive.

The initial coding suggested by Sawtooth Software added no additional information when the respondent said "Some are Very Important, Some are Not". While developing the presentation for the 2010 Sawtooth conference this coding was seen as incomplete. Later in this paper we will show why this incomplete coding should not be used.

The more complete coding was suggested by Paul Johnson of Western Watts. This modifies the initial coding of the Best and Worst tasks. Using the same example, we would alter the initial Best task to the following:

a1	a2	a3	a4	a5	a6	a7	a8
1	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0

This is the same as the "None are Very Important", except that the winner will be the actual best attribute (rather than the zero vector). The idea here is that we know some of the attributes are very important, which means that the attribute selected beats the zero vector. That is the additional information added here.

We also need to modify the Worst task in the same way:

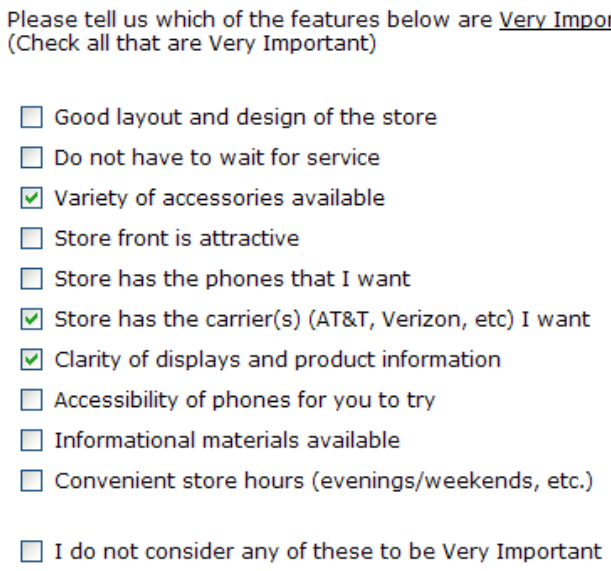
a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	0	0	0	0
0	0	0	0	0	-1	0	0
0	0	0	0	0	0	0	-1
0	0	0	0	0	0	0	0

The winner will be the actual worst attribute. This means the negative of the worst attribute beats the zero vector – suggesting the worst attribute is less than zero.

In summary the revised coding tells us that the best attribute beats the zero vector, while the worst attribute loses to the zero vector. The other attributes we still know nothing about whether they are positive or negative. This additional information is imperative to properly anchor the MaxDiff utilities. While the revised coding provides much more information, it should be noted that we may not gather threshold information about some attributes. If each time an attribute appears it is neither best nor worst, and if the follow up is "Some are, Some are not", then we know nothing about whether the attribute is positive or negative.

III. Anchoring Technique Two: Direct Method

The indirect method requires a follow up question with each MaxDiff task. In addition we also may not gather information about whether some of the attributes are positive or negative. This leads us to consider another technique, which simply asks the respondent to check whether each attribute is above or below the threshold. An example of this direct method is:



This question may be asked after all the MaxDiff tasks. This means no break in the continuity of MaxDiff tasks, and less time than the indirect dual response method. Perhaps most importantly, we get information about whether each attribute is above or below a threshold.

The coding used in this paper involved adding two tasks for each respondent: one representing the attributes above the threshold, and one for the attributes below the threshold. To illustrate this coding, assume there are 8 attributes, and that attributes 1, 3, 6, 7, and 8 meet the threshold of "Very Important". Then we add the following task:

a1	a2	a3	a4	a5	a6	a7	a8
-1	0	0	0	0	0	0	0
0	0	-1	0	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	-1
0	0	0	0	0	0	0	0

The zero vector (last row) wins, meaning that the negations of the utilities lose to zero. Again we have no reference level. The remaining attributes do not meet the threshold and are coded with positive ones losing to the zero vector:

a1	a2	a3	a4	a5	a6	a7	a8
0	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	0	0	0

Adding these simple two tasks informs the model whether each attribute should be positive (meets threshold) or negative (does not meet threshold). Of course if all of the attributes lie on the same side of the threshold then only one task would be added.

Alternative codings were tested, including the binary version where each attribute was compared with a zero vector. This resulted in slightly different utilities, primarily increasing their variance.

IV. Results – Timing and Satisfaction

563 respondents did the direct method only, while 569 respondents did the indirect dual response augment after each MaxDiff task, followed by the direct augment after all MaxDiff tasks.

Comparing these two groups, the direct method is much quicker:

1. Respondents took an average of 4.3 seconds per task to complete the indirect dual response question. So with 15 tasks, the total time is 1 minute 21 seconds. This time computation includes removing 10% of outlier respondents who took more than 40 seconds per task.
2. In comparison the 20 attribute grid with 10 per screen took about 19 seconds of total time.

Given the additional time of the indirect augment, coupled with the dual response break in continuity, we expected respondents to be less satisfied with the survey when they were asked the indirect dual response augment. However, we did not observe any significant change in satisfaction with the survey. On a typical five-point satisfaction scale, the Direct Method shows a slightly higher mean satisfaction of 4.08 vs. 4.00, and a 76% top 2 box score vs. 74% for the Indirect augment.

Overview of Sampling Split - 4 Primary Cells

To compare the two methods most directly, we will focus on the 569 respondents who completed both the indirect dual augment and the direct method. These respondents were assigned to one of the following four cells:

Group	N Size	Attributes	Items Per Task	MaxDiff Tasks
1	163	All 20	4	15
2	129	All 20	5	12
3	142	Better 12	4	9
4	135	Worst 12	4	9

Group 1 will be used for initial comparison and is our baseline. Group 2 is like group 1, but 5 attributes were shown at a time. Group 3 and 4 split will be compared with group 1 to see how well a subset of attributes matches the entire attribute list (more on this later).

We also showed 563 respondents the direct method only. This was done to see if the indirect dual response augment had any measureable impact on the direct results. It did not. So in order to compare the methods in the most direct fashion we will focus on these 4 cells above where respondents completed both methods.

V. Results – Group 1 Baseline

A. Observed Patterns of Choices

Among the 163 respondents of Group 1, we observed the following general patterns:

- 17% of respondents always choose a Mix (some important/some not)
- 72% of respondents use "All Very Important" at least once
    - 61% use at least twice
    - 48% use at least thrice
- 30% of respondents use "None Very Important" at least once
    - 17% use at least twice
    - 13% use at least thrice

So respondents are clearly using the different options in the dual response, sometimes choosing a mix, and other times an "All" or "None". But if we consider all of the tasks across all of the respondents we get the following breakdown of clicks:

All Very Important	**22.2%**
None Very Important	**5.9%**
Mix	**71.8%**

One can see that a Mix is clearly the most common click. This is in line with theoretical expectations. Showing four attributes at a time we should expect the Mix response about 30% to 90% of the time, depending upon how many attributes are above and below the threshold. The more evenly the attributes are distributed, the more Mix responses we expect, as the table below shows.

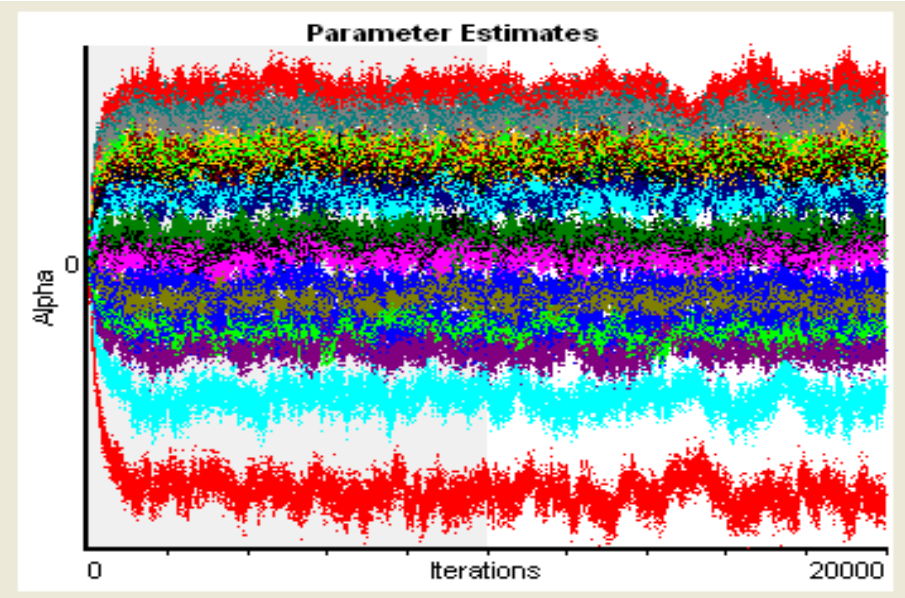
	Show 4 Attributes at a Time			**Show 5 Attributes at a Time**		
Percent Attributes Meeting Threshold	**Prob All > Threshold**	**Prob None> Threshold**	**Prob Mix**	**Prob All > Threshold**	**Prob None> Threshold**	**Prob Mix**
10%	0.0%	65.6%	**34.4%**	0.0%	59.0%	**41.0%**
20%	0.2%	41.0%	**58.9%**	0.0%	32.8%	**67.2%**
30%	0.8%	24.0%	**75.2%**	0.2%	16.8%	**83.0%**
40%	2.6%	13.0%	**84.5%**	1.0%	7.8%	**91.2%**
50%	6.3%	6.3%	**87.5%**	3.1%	3.1%	**93.8%**
60%	13.0%	2.6%	**84.5%**	7.8%	1.0%	**91.2%**
70%	24.0%	0.8%	**75.2%**	16.8%	0.2%	**83.0%**
80%	41.0%	0.2%	**58.9%**	32.8%	0.0%	**67.2%**
90%	65.6%	0.0%	**34.4%**	59.0%	0.0%	**41.0%**

In our case study, Group 2 with 5 attributes showed more Mix responses (79%), again as one would expect. Given the prevalence of Mix responses, it is clearly very important how one codes this information.

B. Convergence in HB

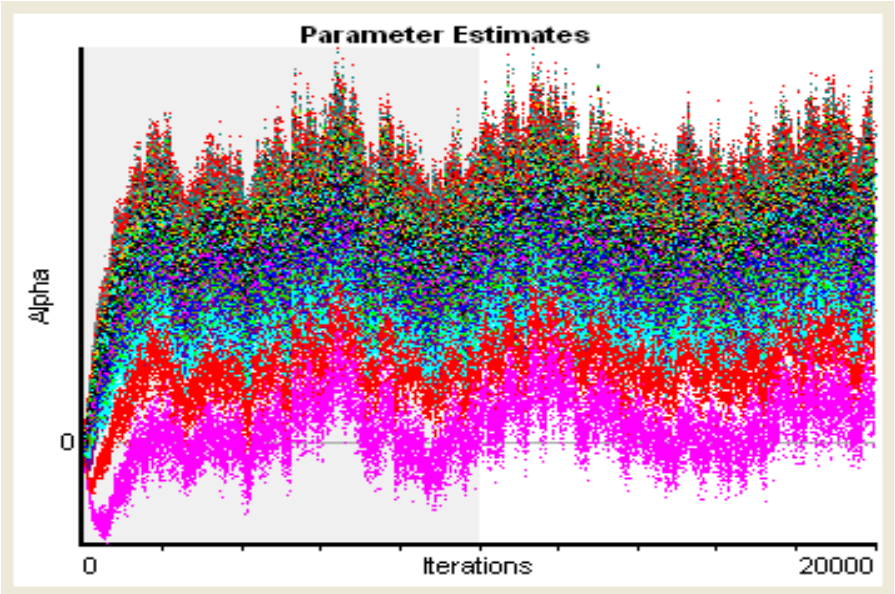
We estimated the utilities using Sawtooth Software's HB CBC, with a prior variance of 1. We first ran the normal MaxDiff utilities without any of the anchoring information. The utilities converged very nicely.

Only MaxDiff Questions



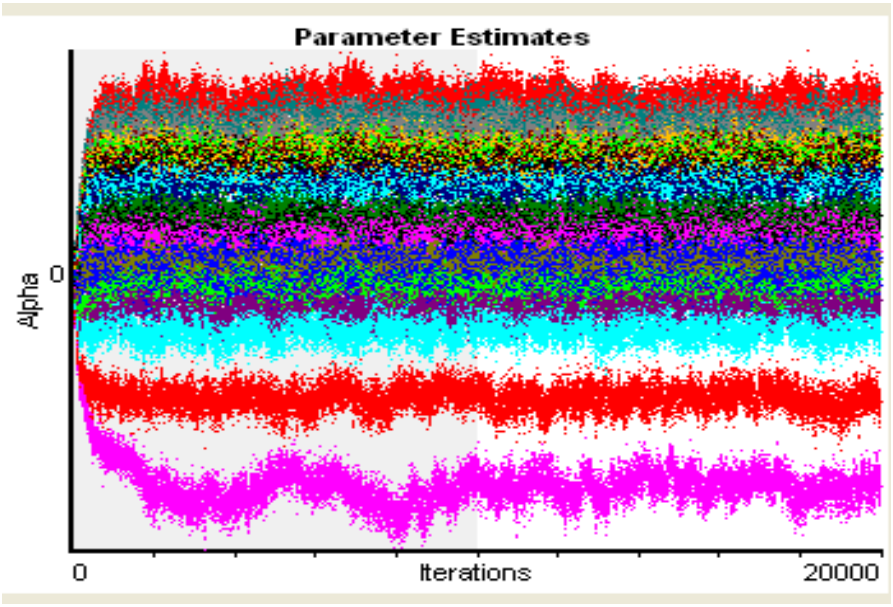
We then added the dual response augment. First we looked at the incomplete coding, where the "mix" response is not coded at all.

Indirect Dual Response Added – Incomplete Coding



As one can see, this did not converge. Playing with the degrees of freedom, prior variance, and number of iterations did not help with convergence. In comparison, when we coded the mixed responses using the revised coding of best and worst tasks, we once again got very nice convergence:

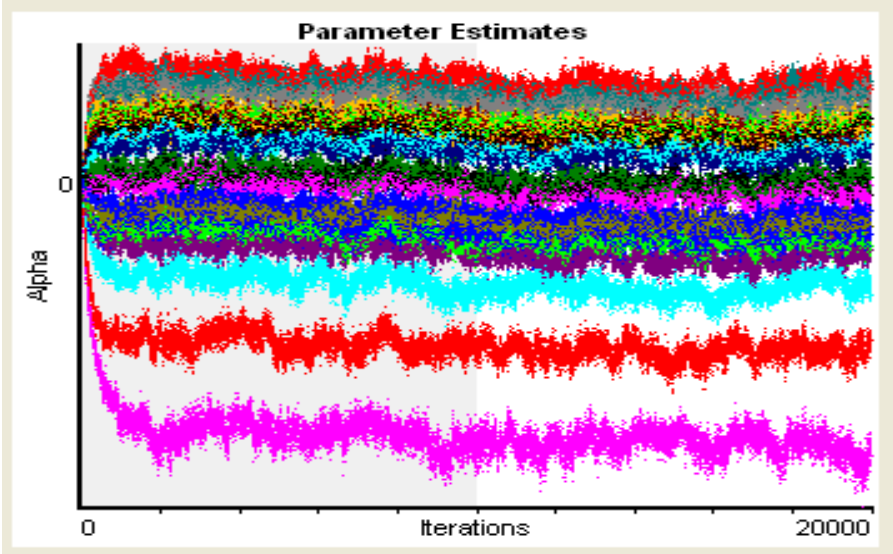
Indirect Dual Response Added – Complete Coding



This alone gave us good reason to implement the coding of the mixed response over no coding of the response. As we will see later, the incomplete coding really should not be used for many additional reasons as well.

Finally, we checked the Direct method where we asked respondents to check the attributes that were "Very Important". This also converged very nicely:

Direct Method



C. Utility Comparison

At the respondent level, the relative utilities from all four methods are nearly identical. If you rank the utilities and run a correlation between the ranks (at the respondent level), one gets an average correlation of .988 with the simple MaxDiff. So all the methods are preserving rank order of utilities. So although we included two holdout tasks with rankings, there was no difference in the ability to predict the rankings.

While the relative utilities of the methods are nearly identical, the absolute utilities are very different. The most important point is that the incomplete coding of the indirect dual response was a complete failure. This again is where we added no information at all when the respondent chose a "Mixed" response. To show just how badly this method failed consider the following table:

Indirect Augment	N	True Expect	Match	Comment
Always Positive	10	All Positive	10	
Always Negative	0	All Negative	0	
Always Mix (No Information)	28	Some Positive	16	10 all +, 2 all neg
Positive and Mixed Only	77		0	All 77 Positive
Negative and Mixed Only	17	Some Negative	0	Al 17 Negative
Positive and Negative (Opt Mix)	31		24	6 all +, 1 all neg
Total	163		50	

If the respondent thinks all the attributes are Very Important then the respondent will always give the positive response in the dual response, stating that all the attributes are very important. We see this happens for 10 respondents (first row of table), and the HB utilities match – giving all positive utilities. On the flip side, respondents who think none of the attributes are Very Important would always give the negative response to the dual response. There are no respondents in this group (2nd row of table) and the HB utilities also reflect that. So far, so good.

But in any other scenario we expect there to be some utilities for a respondent which are positive and some that are negative, reflecting that some attributes are Very Important while others are not. But in fact we rarely see this at all. In these cases, the lack of coding for a mixed response gives no information, and the attributes tend to inherent the non-mixed response from the respondent or the group response. For the 77 respondents who gave a positive and mixed dual response, all 77 had all positive utilities. From the standpoint of information in the model this is consistent, because the model is only seeing a few tasks which are stated to be all positive. The other tasks with a mixed dual response contain no information, which is consistent with a lower positive utility.

In total only 50 out of 163, or 30.7% of respondents have the correct utility structure of all positive/all negative/ or a mix of positive and negative. So we are not getting the anchoring right for the vast majority of respondents.

When we add the coding for the mix response, the results improve dramatically:

Indirect Augment	N	True Expect	Match	Comment
Always Positive	10	All Positive	10	
Always Negative	0	All Negative	0	
Always Mix (No Information)	28	Some Positive	28	
Positive and Mixed Only	77		77	
Negative and Mixed Only	17	Some Negative	17	
Positive and Negative (Opt Mix)	31		31	1 All Negative
Total	163		162	

All but one respondent is consistent in their utility structure of all positive/all negative/ or mix of positive and negative. This one exception was due to respondent inconsistency, where the respondent gave the same attributes an "All Positive" and "All Negative" response.

The Direct method matched the sign structure for all but 3 respondents (160 out of 163). These 3 exceptions were due to inconsistency in the respondent's choices, where the respondent said an attribute was Very Important but it lost to another attribute that was not Very Important.

The clear conclusion is that the incomplete coding of the indirect method is highly inadequate in capturing the mix of positive and negative utilities, where the other methods are extremely successful.

D. Simulated Data Comparisons

Using simulated data, we can show the incomplete coding of the indirect dual response to perform miserably, and that the results get worse as the number of Mixed responses increase. At this point however, we will no longer discuss the incomplete coding as we believe our discussion is sufficient to show it is completely inadequate.

Simulated data also shows that the Direct method is better than the Indirect Dual Response (complete coding). The reason for this is that Indirect method, even with the complete coding may still be indeterminate for some attributes. To better understand this, consider that each attribute is seen a certain number of times per respondent (for example 3 times). Each of those times, the follow up response could be the mixed response. If the attribute is not chosen as best or worst in any of those 3 scenarios, then we have no information about that attribute. This indeterminacy of the attributes increases with the number of attributes shown per task, and as the attributes are more evenly distributed (50% of attributes are positive and 50% negative). For this reason, we do not recommend the indirect method when there are 6 or more attributes shown per MaxDiff task.

The Direct method works extremely well with simulated data, outperforming the Indirect method in almost every set of simulated data. The only case in which the Direct method performs more poorly than the Indirect is when the true utilities of a respondent have small differences relative to the error.

Conclusion here is that in theory the direct method works best. The question is whether real people respond to the indirect augment more accurately than a list of attributes.

VI. Results – Group 3 and 4 Preservation

Group 3 was just like Group 1, but Group 3 saw only 12 attributes. Our initial intent was that Group 3 would have the top 12 attributes, but our initial estimate (based on a sample of 10) was wrong. Group 4 saw a different set of 12 attributes. Groups 3 and 4 had the minimal overlap of 4 attributes.

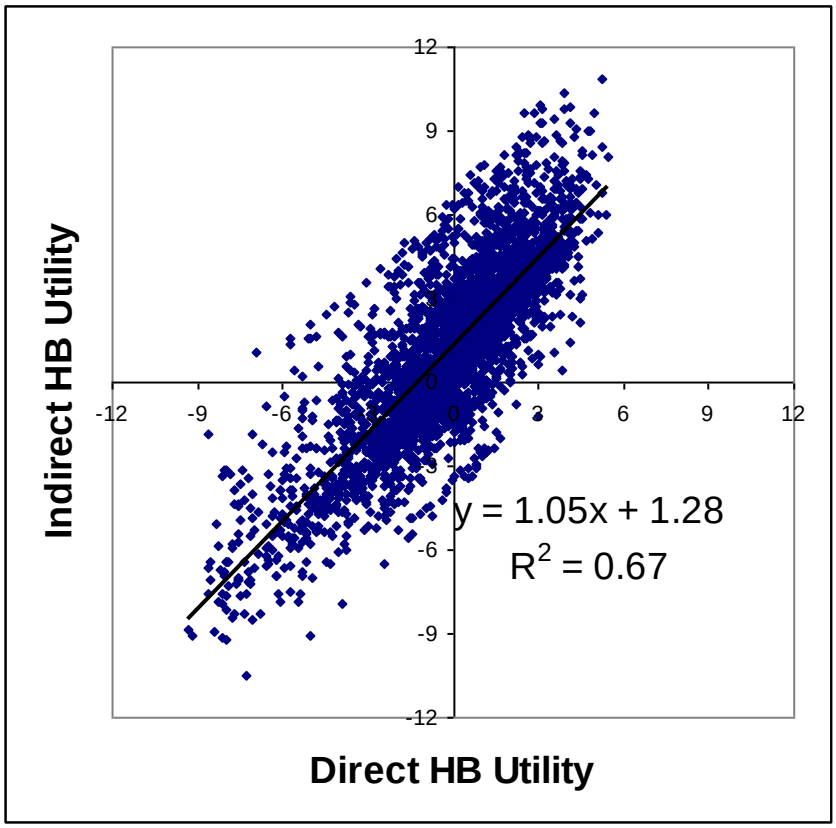
The objective in showing subsets was to test what happened when the anchoring went from all 20 attributes to a subset of 12. In theory, the anchoring should be the same whether respondents saw 12 attributes or 20. In practice, respondents are known to contextualize their responses, and indeed this is what we observed here.

First we noticed that respondents doing the direct approach (which showed 10 attributes on a screen twice), were more critical. That is, these respondents were less likely to say an attribute was Very Important. The table below shows that only 4-5% of respondents clicked an attribute as Very Important in the Direct method, but did not say it was Very Important in the Indirect method. In contrast, about 18-20% of respondents said an attribute was Very Important in the Indirect method but did check it as important in the Direct method. So the check marks definitely indicate a more critical attitude for the Direct approach, at least when 10 attributes are shown per screen.

Repercentaged

Direct Grid	Indirect Grid	4 Att MD/ 10 per Grid	5 Att MD/ 10 per Grid	4 Att MD/ 10 per Grid	5 Att MD/ 10 per Grid
Match Sign		64.60%	60.50%	75.80%	76.70%
Positive	Negative	3.60%	3.70%	4.20%	4.70%
Negative	Positive	17.00%	14.70%	19.90%	18.60%
Pos or Neg	No Info/ Inconsistent	14.80%	21.20%		

This more critical attitude in the Direct method toward which attributes are Very Important is confirmed with the scatterplot of the utilities. In the scatterplot below, each point is the utility for a specific respondent on a specific attribute, showing the utilities from both methods.

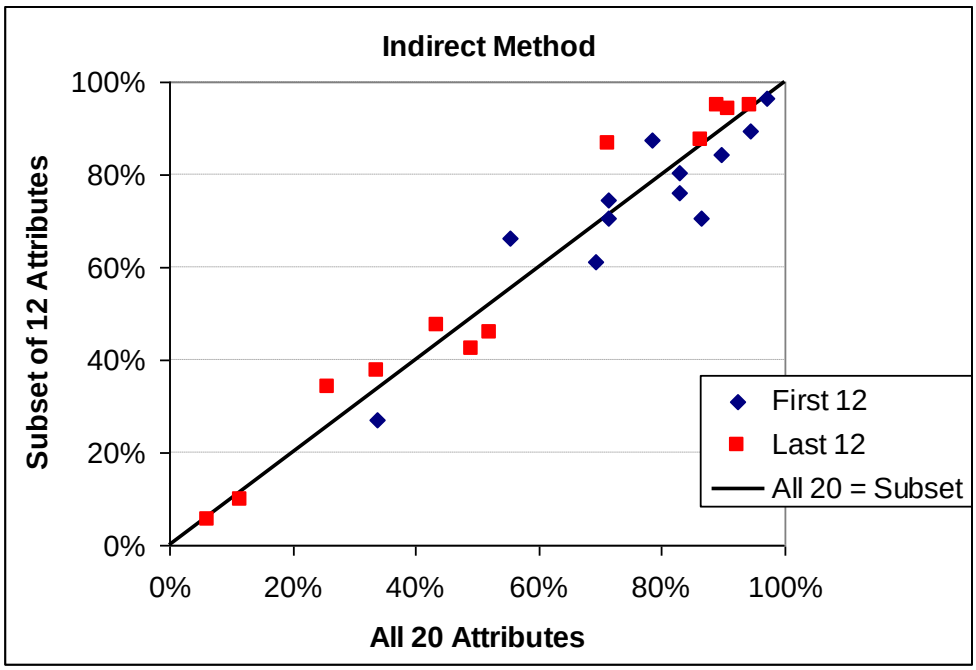


Utilities for the Indirect Dual Response Augment are shifted more positively. So we see that there is a difference between the two, but why?

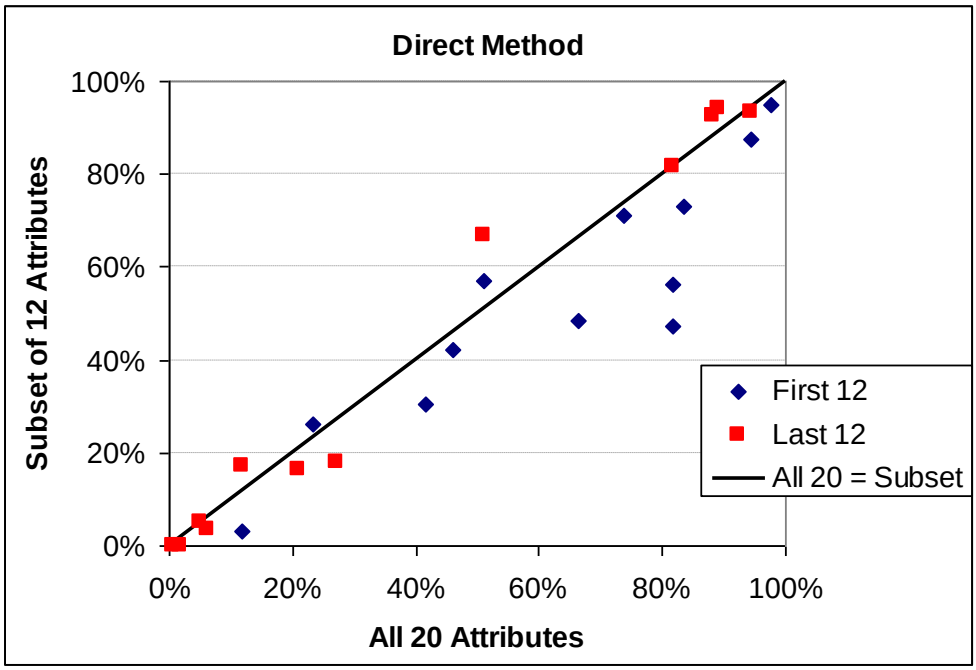
One potential explanation for this is that respondents who see 10 attributes on a screen are comparing all 10 attributes to each other and their Very Important grade is based on these that are Very Important compared to the others. In contrast, with the indirect augment respondents only saw four or five attributes on a screen, and were doing less comparative work to assess whether an attribute was Very Important.

This context sensitive explanation becomes even more plausible when we consider Groups 3 and 4, where only 12 of the 20 attributes were shown. If respondents did not apply contextual relativity then we would expect the two 12 attribute subgroups to be similar to the results from when all 20 attributes are shown.

The scatterplot below shows the percentage of positive utilities for an attribute using the Indirect Method. The x-axis shows the percentage positive for Group 1 doing all 20 attributes. The y-axis shows the percent positive for Groups 3 and 4, who did a subset of 12 attributes. Ideally we would expect all the attributes to fall on or near the line, indicating the same percentage of positive utilities for an attribute whether all 20 were shown or just a subset of 12.



In contrast when we look at the Direct method we see more divergence from the diagonal line.



So the indirect method shows better preservation of the Very Important threshold when only a subset of attributes are shown. This means if one wants to adopt the method that is most likely to capture the absolute threshold one should use the Indirect Dual Response augment. The direct method introduces some contextual relativity – and will change more as the attributes change.

VII. Conclusion

The Indirect Dual Response Method will be indeterminately anchored for some attributes. This indeterminacy is excessive when the incomplete coding is used, and we showed how this led to completely unacceptable results. But even with the revised complete coding of the Indirect method, some indeterminacy occurs. This indeterminacy increases with the number of attributes shown per MaxDiff task, and as the threshold is more evenly distributed (50% of attributes are positive and 50% negative). For these reasons we recommend showing four attributes at a time with the Indirect method, and certainly no more than five attributes at a time. If one must show six or more attributes per MaxDiff task then we recommend the Direct method.

While the Direct method is more accurate in theory, real respondents tend to apply a contextual relativity in evaluating whether an attribute meets a threshold like "Very Important". If one can live with some degree of contextual relativity, then the Direct method is preferable. But if it is important to avoid this contextual relativity for the anchoring then one must weigh the importance of less context dependence against the indeterminacy of the Indirect method.

DIRECTING PRODUCT IMPROVEMENTS FROM CONSUMER SENSORY EVALUATIONS

KAREN BUROS
RADIUS GLOBAL MARKET RESEARCH

ABSTRACT:

Using consumer evaluations to guide product development is problematic when a product fails to achieve its goals. This paper investigates an alternative to Penalty Analysis—sensory drivers and simulation—to understand which attributes play a greater role in driving consumer product acceptance. A key problem with penalty analysis is that it often produces conflicting recommendations for further development. Earlier attempts to simulate changes in product perceptions allowed the researcher too much leeway in simulating changes in product features permitting modifications contrary to respondents' held beliefs about the relationships among the product characteristics. For example, analysis could permit the analyst to simulate the change to product desirability due to simultaneously increasing both its perception of sweetness and tartness. Such a product change would seem impossible to achieve. This paper proposes an approach where product perceptions on attributes were first submitted to factor analysis to obtain a reduced set of factors that captured the underlying correlation structure among individual attributes. Then, the resulting simulator allowed the analyst to make changes to the overall factors rather than the original attributes. This approach more correctly avoided impossible formulation modifications, keeping the product recommendations better in line with the relationship of variables as perceived by respondents.

SETTING THE STAGE:

Consumer product evaluations, conducted through in-home use testing or central location product trial, are a mainstay of product development in the realm of consumer package goods. While a product is submitted to many stages of evaluation during the development process, including evaluations by expert panels, product acceptance by 'real life' consumers is essential prior to product launch.

When a product successfully meets established norms for the company all is well. When it fails to achieve the goal, the researcher is tasked with providing guidance for product modification. Was the product too sweet or too tart? Was the color too light? Was the texture too thick? Was the aroma strong enough? Often consumers are queried for their perceptions on these attributes using a scale similar to the following:

Would say that the texture of the product you tried was…

Much too heavy

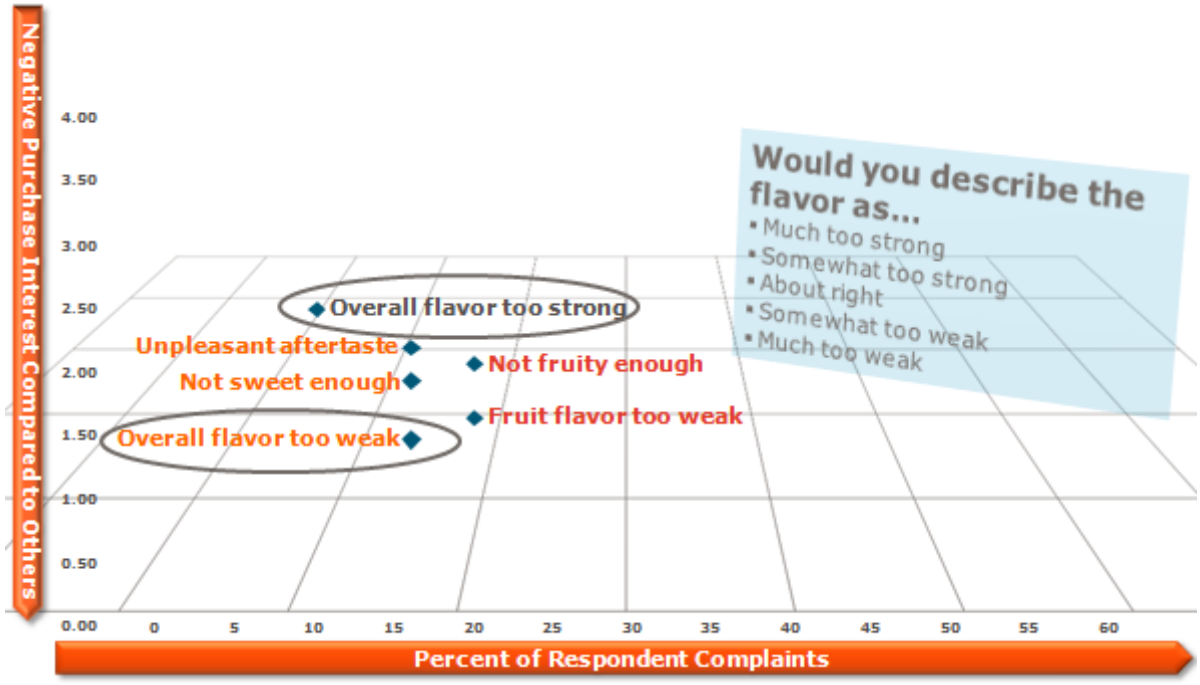
A little too heavy

Just right

A little too light

Much too light

Penalty Analysis is an approach to providing guidance for further development by examining the stated propensity to purchase the product among those consumers rating the product "too heavy" vs. "too light" using the above example. The following chart illustrates this approach to the analysis.



The left axis defines stated propensity to purchase. The horizontal axis depicts the percentage of respondents expressing the complaint. In this example, the 'flavor' is depicted as both 'too strong' by some and 'too weak' by others, pointing to several issues in this analytic approach:

1. We see only an interpretive relationship between product perception and expressed interest in buying the product.

2. We see not only differing product perceptions but likely differing 'tastes' in an ideal product, some may prefer sweeter products than others. Some may prefer their product more full-bodied than others.

3. We are not accounting for those who are evaluating the product as 'just right' in recommending product modifications. If we make the product 'sweeter' would we alienate not only those who consider it too sweet as it stands, but also those who consider it 'just right'?

Perceptual Mapping Linking Product Perceptions to Preferences

In 1986 Sawtooth Software introduced APM (Adaptive Perceptual Mapping) based on Discriminant Function analysis as a tool for the researcher to simulate changes in consumer perceptions of product characteristics. In 1998 Sawtooth Software introduced CPM (Composite Perceptual Mapping) to incorporate both product perceptions and preferences. (See "Mapping Product Perceptions and Preferences", Richard M Johnson, Sawtooth Software, Inc., 1998.)

The goal of the CPM approach, in linking perceptions on product characteristics to preferences, is the give the researcher greater ability to recommend how changing product perceptions could increase demand for that product.

While both of these approaches can produce insights for the marketer, neither is directly applicable to the issues of product testing through consumer evaluations.

Both mapping procedures involve the determination of a consumer's 'ideal point' in product perceptions. That is, how sweet, or tart or dark does the consumer want that product to be. This is an important consideration that will be discussed later.

The Realities of Product Testing:

There are many realities in the world of product testing that impact the ability to predict the effect of changing product perceptions on preference. A few are noted here:

- Typically consumers evaluate one to four products either in-home or in central location, a very small number of observations for modeling purposes.
- These are consumers, not experts, who will use scales to their own liking and will seek differing product characteristics.
- There is a great deal of multicolinearity across the attributes defining the perceptions.
- Alterations to a product on one dimension, e.g., sweetness, will affect others, e.g., tartness.

Keeping these realities in mind, any modeling to estimate changes in preference should be considered directional, providing guidance to the research and development team for future product modifications. Further, while the goal of this effort is to guide development of the products being evaluated, this effort can also assist in identifying flanking products that might satisfy differing desired 'tastes'.

The Conceptual Approach:

Conceptually, the approach is straightforward. In practice it is problematic.

The core idea is to establish the relationship between attribute evaluations and purchase intent for each respondent. Then take those relationships into an Excel-based simulator in such a way that the researcher can increase or decrease respondents' stated perceptions of the product, reading the resulting change in purchase intent.

The first step is to establish a baseline perceptual reading of the attributes in question, e.g., how sweet is the product they tasted? A simple bi-polar scale can do this, as illustrated below:

Not at all Sweet ____ ____ ____ ____ ____ ____ ____ ____ ____ ____ Very Sweet

Purchase intent is measured on a standard five-point Likert scale, although any multi-point scale could be used:

Definitely would buy

Probably would buy

Might or might not buy

Probably would not buy

Definitely would not buy

Although not implemented in the work thus far, a depiction of the respondent's 'ideal' product on the bi-polar scales would be highly desirable, as will be discussed in subsequent sections.

An Early Model:

With multiple (2 to 3) product evaluations for each respondent, Latent Class regression was first used to determine the likely influence of each rating on purchase intent. Simply described, Latent Class regression estimates individual level coefficients and for each attribute and intercepts. Bringing these into Excel, one can estimate the probability that the respondent belongs in each rating point of the five point ordinal scale.

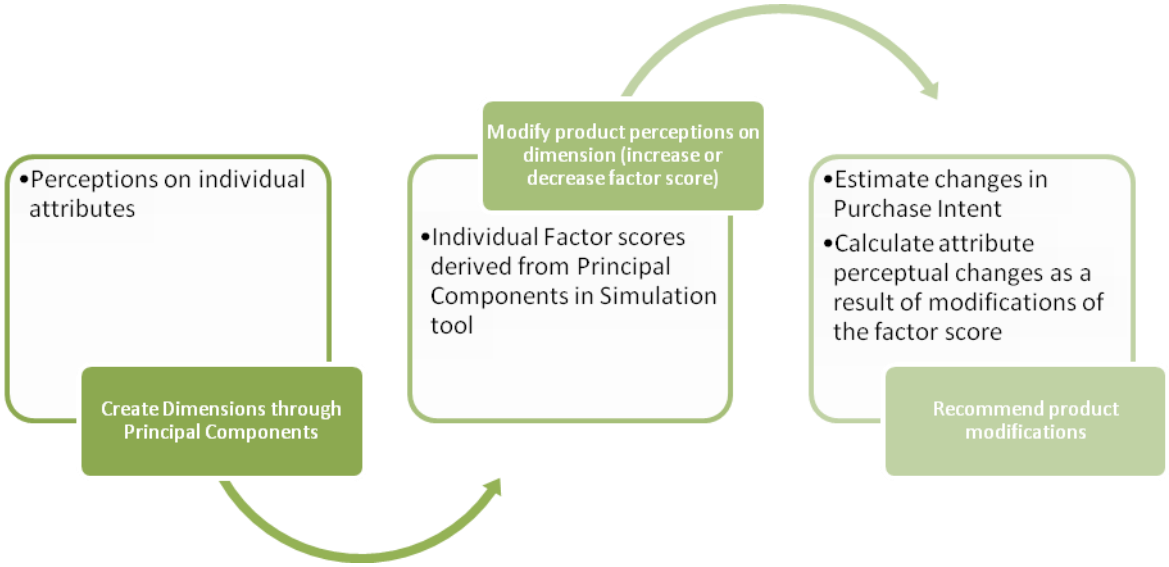
In simulation, the respondent's ratings can be increased/decreased by one or more rating points on the bi-polar scale from the current position, keeping in mind that the scale is bounded at each end, that is, ratings cannot be shifted beyond the ends of the scale.

Under this approach, all respondents are moved simultaneously under the reasoning that, if the product is made sweeter, for example, that will result in a positive shift in purchase intent for respondents having positive coefficients and a lessened purchase intent for respondents having negative coefficients.

There are some obvious problems with this approach, the most major of which is multicolinearity across the attributes studied. A second major shortcoming is that the user of the simulator can shift several attributes simultaneously and arbitrarily, often defying logic.

Building in Multicolinearity:

The key 'next step' was to build a simulator that would address the multicolinearity problem by allowing the researcher to modify the product along a dimension spanning multiple, correlated attributes. In this way, simultaneous modification of multiple attributes could be incorporated. Diagrammatically, the flow is:



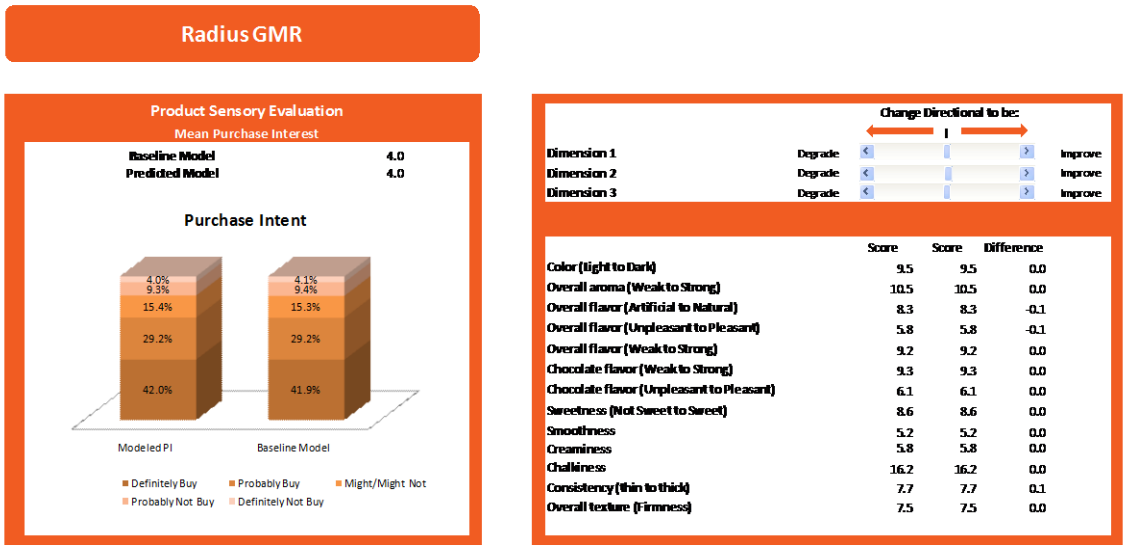
For this purpose, Principal Components Factor Analysis using a Varimax rotation with Kaiser Normalization was used (run in SPSS). An example of this output (Rotated Component Matrix) is shown here:

Rotated Component Matrixa

	Component		
	1	2	3
Pleasant Flavor	.882	.228	.087
Pleasing Chocolate	.842	.273	.087
Natural Flavor	.840	.140	.082
Creaminess	.639	-.079	.544
Smoothness	.631	-.178	.461
Chalkiness	-.594	.191	-.106
Strong Flavor	.223	.838	.079
Strong Chocolate	.342	.769	.105
Overall aroma	-.035	.750	.114
Sweetness	.029	.538	.240
Color of product	-.173	.416	-.070
Overall texture	.121	.201	.838
Consistency	.155	.194	.828

In this example, three scores are derived, the first relating to ‘pleasant flavor’, the second depicting ‘strong flavor and aroma’ and the third relating to ‘texture and consistency’. It should be noted that some attributes, such as ‘creaminess’ loaded on both the first and third dimensions. A Varimax rotation was chosen so that the dimensions would be ‘independent’.

A screenshot of the simulator produced in this work is shown here:

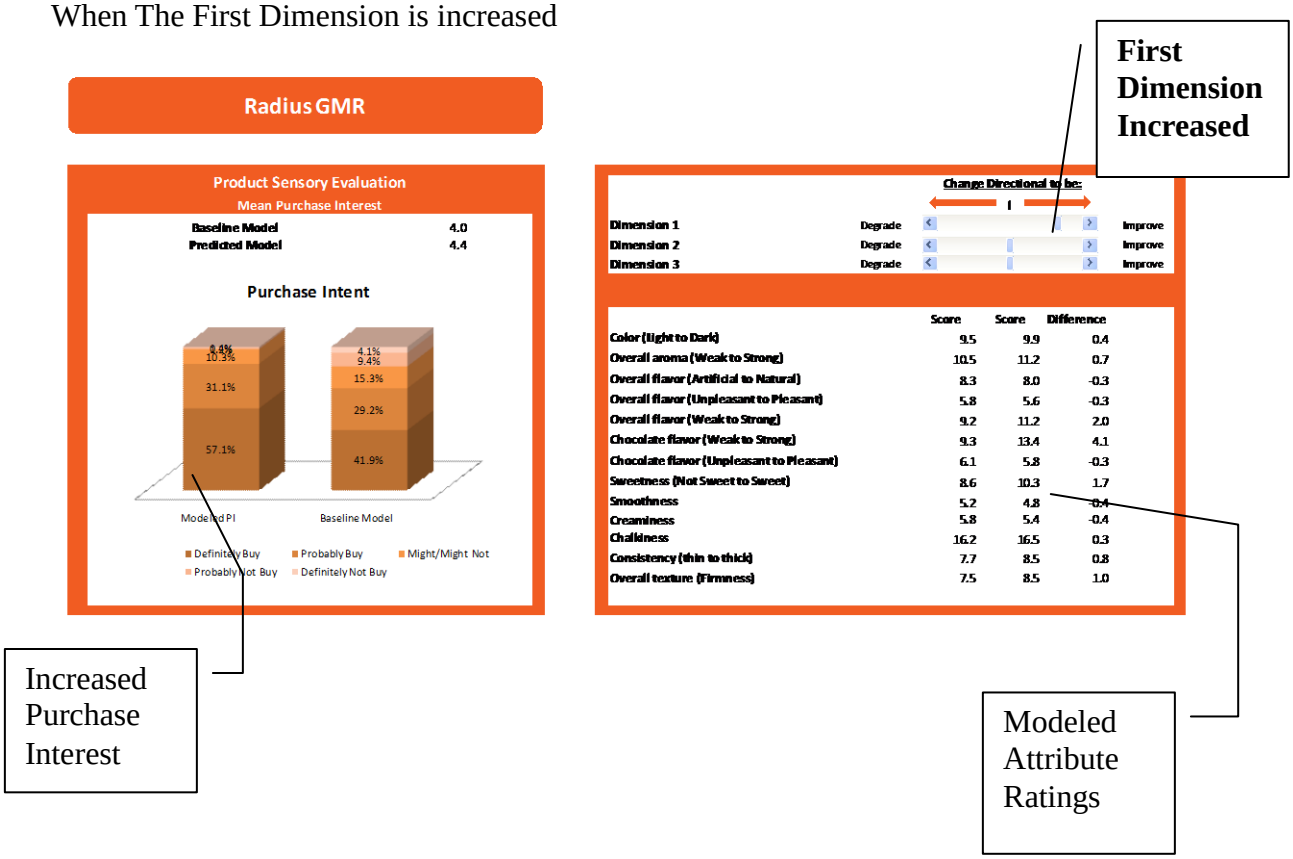


On the left the ‘baseline’ purchase interest and the ‘modeled’ purchase interest percentages are shown across the five point purchase interest scale. On the top right, three sliding scales, representing each of the factor scores, are shown. The researcher can increase or decrease the factor scores for each respondent by sliding the marker to the left or right, resulting in a modeled purchase interest score depicted on the left. The altered factor score is re-translated back to attribute scores bounded by the scale used. In other words, once the modeled respondent level score reaches the upper or lower bounds of the scale, it cannot be further modified. The original scores and the modeled scores are shown on the lower right.

There is a serious question regarding the appropriateness of a Principal components factor approach to form the dimensions due to the covariance structure underlying the factors. To better understand this question, on this set of data, additional principal components factor analyses were performed on the attribute scores resulting from the dimensional alterations in the factor scores through modeling. While this does not resolve the issue, the findings are informative.

In the tables below are shown the Rotated Component Matrices under three situations – the original matrix, the matrix when the first dimension is increased and when the second dimension is increased. First is shown the simulator screenshot indicating the movement in the first dimension and the resulting attribute and purchase intent changes.

When The First Dimension is increased



Similar movement was conducted for the second dimension, returning the first to a neutral position.

In all cases, a three factor solution emerged as the solution above an eigenvalue of 1. The three Rotated Component Matrices are very similar in structure:

Original Rating Rotated Component Matrix

	Component		
	1	2	3
Pleasant Flavor	.882	.228	.087
Pleasing Chocolate	.842	.273	.087
Natural Flavor	.840	.140	.082
Creaminess	.639	-.079	.544
Smoothness	.631	-.178	.461
Chalkiness	-.594	.191	-.106
Strong Flavor	.223	.838	.079
Strong Chocolate	.342	.769	.105
Overall aroma	-.035	.750	.114
Sweetness	.029	.538	.240
Color of product	-.173	.416	-.070
Overall texture	.121	.201	.838
Consistency	.155	.194	.828

Dimension 1 Increased Component Matrix

	Component		
	1	2	3
Pleasant Flavor	.841	.348	.065
Pleasing Chocolate	.806	.361	.080
Natural Flavor	.788	.285	.076
Creaminess	.640	.093	.550
Smoothness	.629	.025	.507
Chalkiness	-.599	.061	-.195
Strong Flavor	.287	.832	.101
Strong Chocolate	.267	.824	.169
Overall aroma	.107	.755	.099
Sweetness	.210	.629	.158
Color of product	-.141	.513	.313
Consistency	.173	.284	.808
Overall texture	.208	.257	.799

Dimension 2 Increased Component Matrix

	Component		
	1	2	3
Pleasant Flavor	.899	.147	.126
Natural Flavor	.851	.064	.070
Pleasing Chocolate	.850	.229	.144
Chalkiness	-.539	.181	-.240
Strong Flavor	.186	.837	.114
Strong Chocolate	.278	.764	.185
Overall aroma	-.092	.738	.152
Sweetness	.006	.527	.215
Color of product	-.099	.424	-.200
Overall texture	.108	.285	.775
Consistency	.139	.308	.773
Creaminess	.544	-.052	.653
Smoothness	.553	-.153	.567

Even given these similarities, it is worthwhile exploring other approaches to forming the dimensions, such as a Partial Least Squares approach.

THRESHOLDING:

Another problem arises in the current approach that can be addressed via an 'ideal point' analysis. The regression-based approach undertaken in this paper assumes a linear-type relationship between the attributes and purchase interest. This is likely not entirely true. Even though one may want a product to be sweeter, there likely is a point of too much sweetness which must be accounted for.

The 'ideal point' measurement cited earlier, as introduced in Adaptive Perceptual mapping, is one avenue deserving further exploration. The attempt here to 'bound' the scale, restricting it to the points on the bi-polar scale was a first attempt. The author agrees that further work is required in this regard.

A Study of the Diffusion of Alternative Fuel Vehicles: An Agent-Based Modeling Approach[1]

Rosanna Garcia
Northeastern University
Ting Zhang
Xi'an Jiaotong University, (posthumously)

Introduction

In this paper, we focus on the eco-innovation, alternative fuel vehicles (AFVs). As automotive firms invest in improving AFV technologies and manufacturing processes, they, as well as governmental agencies, want to know what policies impact consumers' choices for AFVs. It is essential to understand the conflicting forces that come into play when diffusing AFVs. Consumers want to maximize utility but also minimize costs; manufacturers want to maximize profits, and governmental agencies want to maximize social benefits (air pollution reduction). Extant studies typically have taken a myopic viewpoint of the issue whether it be from a consumers' perspective (e.g., Beggs et al., 1981; Brownstone et al., 1996; Byrne and Polonsky, 2001; Urban et al., 1996), a manufacturer's perspective (Kim et al., 2003; Whitehead, 2001), or a policy perspective (Agrawal and Dill, 2007; Winebrake and Farrell, 1997).

It is the aim of this paper to develop a multi-agent model grounded within empirical data to study the interactions among automobile manufacturers, consumers and policy makers. We focus the study on the impact of three specific mechanisms; governmental policies, word-of-mouth, and technology change on the diffusion of AFVs. Our model combines engineering design optimization (manufacturer perspective) with choice-based conjoint data (consumer perspective) in order to simulate the dynamic marketplace. We use the Michalek et al. (2004) game theoretic model, as the foundation for our study, but expand upon it by building in heterogeneity of consumer demand and competitive influences. We conduct experiments implementing different mechanisms (both government initiated and market initiated) to explore the resulting impact on the diffusion of AFVs within the simulated model.

This paper makes both methodological and empirical contributions. From an empirical perspective, we extend eco-innovation diffusion models by seeking to indentify the factors that are most effective in encouraging the adoption of AFVs. We ground our investigations in consumer behavior theory, consumer choice theory and policy theories, thereby, advancing our knowledge of the interaction between consumer behavior and policy goal setting. From a methodological perspective, we extend upon recent studies (Journal of Business Research special issue2007) by demonstrating how to ground a multi-agent micro-simulation in empirical studies to predict consumer responses to policies. We do this by combining an engineering design optimization model with a consumer choice model. By allowing the interaction among multiple agents, each with unique optimization goals, this paper provides a method for predicting heterogeneous consumer response to automotive design changes.

[1] This conference paper is an earlier version of the paper, Zhang, Gensler, and Garcia, "A Study of the Diffusion of Alternative Fuel Vehicles: An Agent-based Modeling Approach", *Journal of Product Innovation Management, (forthcoming)*. For details on the model, please refer to this publication.

THEORETICAL PERSPECTIVE

In diffusion of technological innovations studies, theory focuses on whether innovations are driven by technological development (technology push) or by demand factors (market pull) (Adner and Levinthal, 2001). Technology push comes when manufacturers put innovations into the marketplace and withdraw existing products. Interestingly, the years 1899 and 1900 were the high point of electric cars in America, as they outsold all other types of cars. Due to technological advances, there is modern day resurgence in 'pushing' EVs back into the market to displace the ICE (Internal Combustion Engine). There are a number of companies that only sell electric vehicles; Norway's Think! and the California-based Tesla are just two.

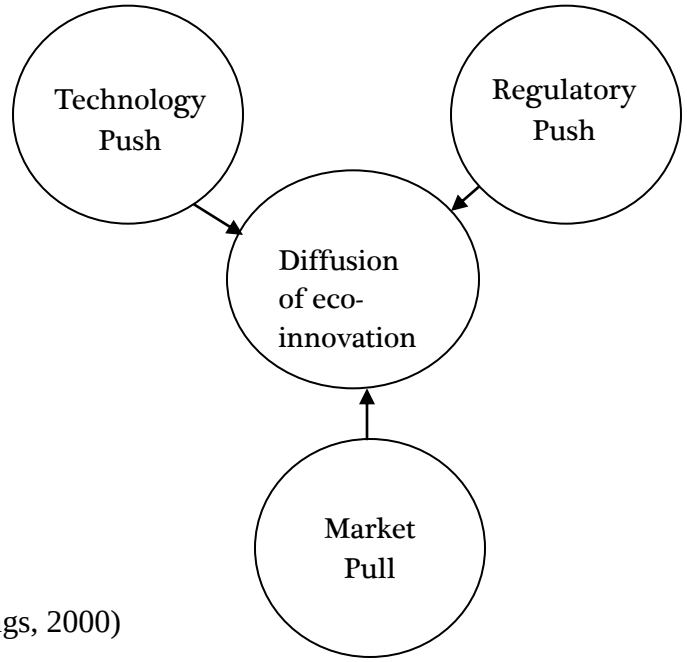
However, existing manufacturers are not willing to take the risk to solely offer AFVs. Introducing AFVs to the market evokes additional cost for product development and manufacturing beyond what is required for traditional products. The premium consumers are willing to pay for an eco-friendly product is small (Loureiro et al., 2002). As long as consumers do not have a strong preference for AFV such a sustainable behavior might lead to a prisoners' dilemma. Technology push is therefore only effective if consumers are willing to buy AFV. We investigate the market situation to determine if a technology push would be effective in the diffusion of AFVs.

Market pull may be another important factor to speed diffusion of AFVs. Today, AFVs are more expensive than environmentally harmful vehicles and come in limited car models. The question is then how to create market pull, i.e. how can consumers' preferences for AFVs be improved? Previous research stresses the importance of consumers' domain-specific knowledge as an antecedent of innovation adoption (Meuter et al., 2000; Moreau et al., 2001). Information about product attributes is often easy to collect. In case of AFVs, consumers can easily collect the information about the car's attributes, for example, on the manufacturer's website or by consulting a dealer. However, to gain information about the actual product or attribute performance is more difficult because it requires experiencing the product (e.g., car reliability, fuel efficiency, driving behavior). It is well understood that consumers rely more on other consumers or experts to assess actual product performance than on company information because they evaluate this information as more reliable (Herr et al., 1991; e.g., Kopalle and Lehmann, 1995). That is also the reason why word-of-mouth is considered as a powerful marketing tool (e.g., Goldenberg et al., 2001a). Many studies show that word-of-mouth positively affects sales (e.g., Chevalier and Mayzlin, 2006; Dellarocas et al., 2007; Duan et al., 2008). When consumers communicate about their product experience, they not only communicate their attitude towards the product but they also transfer product-specific information. The receiver of the message gains new information and might eventually update her knowledge when she considers the information as valuable (Weiss et al., 2008). Hence, word-of-mouth can be an important driver of diffusion of AFVs.

Factors of technology push and market pull are not always effective in diffusing eco-innovation, thus, regulatory intervention is frequently required to facilitate diffusion. Failure by firms and consumers to embrace green or eco-friendly products is known as a problem of 'externality'. Externality is at the center of environmental economics (Baumol and Oates, 1988). It arises when a transaction between parties affects a third party, positively or negatively. For example, a firm might produce and sell a good to a consumer to both the firm's and the consumer's advantage, but the production may negatively impact society. When externality

exists, the market system will typically lead to an inefficient outcome because the impact on any third parties is not considered by the parties participating in the exchange. Due to the externality problem of eco-innovations, government intervention is often required with a regulatory framework (regulatory push or pull) (Rennings, 2000). Thus, to fully explore the mechanisms that can effectively impact the diffusion of AFVs, we must consider three perspectives: (a) regulatory push, (b) technology push, and (c) market pull (see Figure 1).

Figure 1.
Determinants of Eco-Innovations

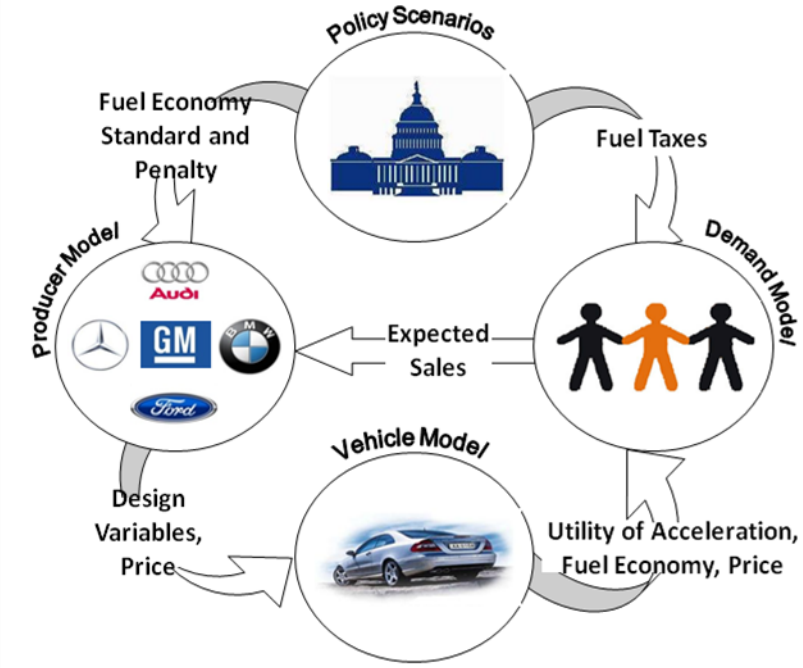


(source: Rennings, 2000)

METHODOLOGY

We use an agent-based model (ABM) for our micro-simulation. Manufacturer agents seek vehicle designs that maximize their profits. Consumer agents seek a vehicle from the available options that maximizes their utilities. A single government agent seeks to encourage the spread of environmentally friendly vehicles by establishing policies. The government agent's behavior can influence the vehicles' design and production behavior as well as consumers' purchasing behaviors, thereby, shifting the marketplace equilibrium. Using an ABM allows us to evaluate the interactions among the different push and pull strategies by the agents. (See Figure 2).

Figure 2.
Interdependencies among Manufacturers, Consumers, and Government



There are numerous reasons we use an ABM. In the ABM, our manufacturer and consumer agents can be modeled as heterogeneous, allowing us to extend upon extant studies that can only consider homogeneity in consumer preferences or singularity in engineering design choices (Michalek et al., 2005; Michalek et al., 2004). Since ABMs can be instantiated with empirical data, we can begin to predict how actual marketplace mechanisms impact manufacturer and consumer choices in speeding diffusion. Additionally, an ABM allows evaluation of the dynamics among agent decisions to identify causality of outcomes.

ABM development

Although we study three types of mechanisms, we identify four types of agents in our model: 'manufacturer agent', 'consumer agent', 'government agent' and a 'vehicle agent'. Manufacturer agents design and produce vehicles; consumer agents purchase vehicles that provide profit to manufacturer agents; the government agent influences both manufacturers and consumers by regulation policies. Creating a 'vehicle' agent allows us to easily include multiple vehicle designs into the model. Manufacturers can choose from thousands of possible vehicles to introduce to the marketplace as price and fuel economy are taken as continuous variables.

Manufacturer Agents. Manufacturers seek to optimize profit. We base the profit function for our manufacturers on previous research (Michalek et al., 2004). This provides us with an extant engineering design optimization model to validate our ABM. Michalek et al. (2004) provide an analytical model of engineering performance, consumer demand and manufacturing costs that are evaluated using game theory to simulate competition among firms to predict design choices at market equilibrium. To account for competition in the design of vehicles, we seek market (Nash) equilibrium for each competing manufacturer. In order to search for the equilibrium point, simulated annealing (Brusco et al., 2002; Černý, 1985; Kirkpatrick et al.,

1983) was employed in which each manufacturer separately optimizes its own profit while competitor manufacturer decisions are held constant.

We use a simulated annealing algorithm to allow us to easily model any number of competitors, not limiting ourselves to 2 player games. For our experiments, we had 6 manufacturers each producing one type of vehicle. We allowed each manufacturer to produce only one type of vehicle to ease verification and validation of the model and for ease of results interpretation. The output from the ABM for each manufacturer is the retail price of the automobile, the miles per gallon/miles between charges and the total profit from sales.

Consumer Agents. The consumer agents are modeled to make purchasing decisions based on their preferences. In our research, we use choice-based conjoint data from an empirical study we conducted to model heterogeneous consumer agents. We, thus, determine the probability of choice of each product based on its utility compared to all other products that are offered. We also consider a 'none' option and we hence do not force the consumers to purchase a vehicle. The utility of consumer i for vehicle j, depends on the consumer's partworths as determined from the products' characteristics v, including body type, fuel type, mpg/mpc, price and the policy, and is the sum of all partworths:

$$u_{ij} = \sum_{v \in V} \beta_{iv} \cdot x_{jv} \qquad (1)$$

In conjunction with AutoWeek, the online newsletter for car aficionados, we collected 7595 responses to our survey. Respondents answered 12 choice tasks with 3 alternative vehicles and a none-option. Further in the survey, respondents were asked how many owners of AFVs they have talked to and 3 questions about their knowledge about AFVs which were measured on a 1-7 scale (fueling, maintenance, and sticker price). The reported number of AFV owners a consumer has talked to, WOM_i, represents the word-of-mouth activity in the market. We assume that word-of-mouth affects consumer preferences for the different attributes. Further, we assume that current product-specific knowledge affects a consumer's preferences. To take these dependencies into account, we model the partworths of the utility function as a function of word-of-mouth and knowledge (Lenk et al., 1996).

The parameter for the covariate WOM is the same across all respondents because dependence is assumed (Lenk et al., 1996). We estimate the parameters by using a hierarchical Bayes (HB) MNL model which uses a continuous representation of heterogeneity, i.e. individual parameters are estimated (Lenk et al., 1996; Rossi et al., 1996). Each consumer agent is initialized with the individual partworths corresponding to an individual survey respondent (Garcia et al., 2007). The partworths reflect the knowledge base of the individual in regards to AFVs. We expect this knowledge to be higher than a more general population because our survey respondents are automobile buffs.

Government Agent. The government agent establishes policies to influence the vehicle production of manufacturers and the purchasing behavior of consumers. Policies can be directed at manufacturers and/or consumers. We investigate the impact of Corporate Average Fuel Economy (CAFE) regulations on the penetration rate of AFVs in our simulated marketplace. The penalty charge for noncompliance with CAFE is ρ =\$55 per vehicle per mpg over the limit

(based on Michalek et al., 2004). In this study, the government agent is exogenous, thus, it is not seeking any optimization functions. We discuss this limitation of the study in the conclusion.

Regulatory push for the diffusion of eco-innovations is a common occurrence around the world. In the United States, regulations establishing a Corporate Average Fuel Economy (CAFE) have been set for over 25 years. CAFE regulations establish minimum average fuel economy standards that each manufacturer's vehicle fleet that sells in the US must meet to avoid penalties. This regulatory push targets manufacturers but the National Highway Traffic Safety Administration also claims that consumers will save $100 billion in fuel costs over the lifetime of vehicles that fall under the rule. These potential cost savings might influence consumers to buy an AFV. We investigate the impact of the current CAFE standards on manufacturers' vehicle design choices and the resulting indirect impact on consumers' choices. We choose to evaluate this regulation because of the trickledown effect from manufacturer to consumer.

Base Case. The base case is 'status quo' as established by the conjoint results. This status quo can be considered the current market situation given only 6 vehicles were available in the marketplace. Using empirical data helps to establishing the base as close as possible to the current reality of the marketplace, which allows us to be predictive in examining the impact of shocks to the system in the form of our experiments.

The manufacturer is initially randomly assigned one of the six types of vehicles to manufacture (gas-sedan, gas-SUV, hybrid-sedan, hybrid-SUV, EV-sedan, EV-SUV). Manufacturers enter the marketplace with their assigned vehicle and compete against other manufacturers. Consumers evaluate the vehicle designs offered by the manufacturers and make a purchase decision by choosing one of the six offerings or choosing the 'none' option. In this way all the manufacturers compete against each other for a limited number of consumers. Manufacturers not satisfied with their profit levels, can re-design the vehicle if they are willing to incur the R&D investment penalty. More than one manufacturer can design the same vehicle. We frequently found that two to three manufacturers would design a vehicle that only differed in miles per gallon and price. For example, in the base case two manufacturers designed gas sedans but one offered the vehicle at 26.14 miles per gallon (MPG) and another offered it at 27.61 MPG.

In the base case, the market reaches equilibrium after about 20 iterations. In equilibrium, no manufacturer is better off by producing a different type of vehicle and no consumer is better off by picking a different type of vehicle. Averaging across the 400 iterations of the base case analysis we found that gasoline engines take the largest market share at 42.49% (34.94% for sedans and 7.55% for SUVs) followed by 33.16% for hybrid sedans (there were no sales of hybrid SUVs). We find in the base case that the 'none' option has 23.9% of market share, thus, signifying that there are other types of vehicles that would provide better utility to consumers, but are not available in the simulated marketplace. In the base case, no manufacturer builds electric sedans and only one electric SUV is ever sold.

The mode for MPG of the 400 iterations is 27.61 for gasoline sedans, 23.35 for Gasoline SUVs, 36.14 for hybrid sedans, and 108.13 for electric SUVs. These are the equilibrium vehicle designs. The average vehicle price is set by each producer, which reflects the manufacturing costs. Each experiment is described below.

EXPERIMENT RESULTS

Following on Figure 1, we conducted three experiments in order to better understand the mechanisms that can speed the diffusion of AFVs:

Implementation of Regulatory Push: CAFE standards.

Implementation of Technology Push: AFV Mandate

Implementation of Market Pull: Word-of-mouth (WOM)

Experiment 1: Regulatory Push - CAFE standards imposed.

Following on previous MUSES studies, the government agent is exogenous to the system. The regulatory goal is to minimize air pollution by encouraging purchases of AFVs through the implementation of CAFE 1973 standards. As previously noted, any vehicle not built to the 27.5 MPG standard must pay a fee. This fee directly impacts the profit level of the manufacturer as described in equation 7. There are little changes in the design of gasoline sedans. This is not surprising because in the base case the most popular type of gas sedan has a fuel efficiency of 27.61 MPG, which already surpasses the 27.5 MPG 1973 CAFE standard. The biggest change we see is that SUVs become more attractive. Hybrid SUVs gain market share as it is now offered at a higher MPG with a lower price than the gasoline SUV (recall in the base case there were no hybrid SUVs produced by manufacturers).

Thus, the overall effect is that we see the market share for AFVs increases because of the introduction of hybrid SUVs. CAFE is successful in increasing the market share of hybrids (sedan and SUVs) by 9.5%, showing that it can be effective in helping with the diffusion of AFVs. However, the social good (air pollution improvement) decreases because market share for the fuel inefficient gasoline and hybrid SUVs increases by more than 35%.

Experiment 2: Technology Push – AFV mandate.

In experiment 2 we look at what would happen in the marketplace if manufacturers only produced AFVs, both hybrid and electric. What is interesting to note in the results is that hybrid sedans take more than 65% of the market share and electric vehicles continue to take less than 1% market share. Our results show that manufacturers introduced different variations of electric sedans but few consumers were willing to buy them. This is likely due to their manufacturing cost, which results in a high retail price (average over $100,000). Hybrid SUVs become popular because this is the only choice for those SUV-loving consumers (5.5% of market share.) What is surprising is that the increase in the share of the ‘none’ option is rather small. These results support the idea that given no other choice, American drivers will be satisfied with hybrid options. Technology push could be an important mechanism for speeding the diffusion of AFVs if the price is affordable.

Experiment 3: Market Pull – WOM considered.

In this experiment we look at the impact of word-of-mouth on the diffusion of AFVs. We model WOM in a fashion similar to Toubia et al. (2009) who quantified social interactions as the number of adopters of an innovation that an individual had spoken to about the innovation. The WOM adjustment factor, obtained from the conjoint results as described in the ‘consumer agent’ model description, is multiplied by the number of reported people spoken to. We found there is a positive influence from WOM on the adoption of electric vehicles both sedans and SUVs. It was

interesting to note that the more expensive the vehicle, the greater was the preference for that vehicle due to word-of-mouth. Because consumers are willing to pay more for AFV due to WOM, manufacturers are able to sell the more expensive vehicle to a niche market despite the higher price (average price now is over $136,000). There was a negative impact from WOM on SUVs, both hybrid and gasoline engines. Because of the negative perception about SUVs, hybrid SUVs were not manufactured and gasoline SUVs lost market share compared to the base case Overall, we find that WOM had a positive impact on the diffusion of AFVs and helped to increase the social good by decreasing the preference for the fuel inefficient SUV (17.9% decrease).

DISCUSSION

The intent of this model is to provide practical insights to governmental policy makers and automotive manufacturers on how consumers' preferences may be altered for AFVs. It also provides a methodology to investigate the diffusion of other types of eco-innovations.

The base case supports the overall market consensus about hybrids. The marketplace in 2010 has seen a movement away from 'mild hybrid drivetrains' because they do not provide the gas-savings that are expected out of hybrids (Edmunds Inside Line).[2] The base case simulation supports this trend by showing that price is not the major issue as much as limited benefits from marginally higher fuel economy are of concern to consumers. Consumers require higher fuel economy in order to adopt hybrid vehicles, and in today's marketplace, the benefits from the higher cost for hybrid technology are marginal.

In regards to electric vehicles, our base case shows there is no interest by consumers. The conjoint results indicate the price is the most important attribute for electric vehicles with fuel economy (miles per charge) as secondary. Survey respondents did not value EVs unless they had over 100 MPC. In the base case simulation the few respondents interested in electric SUVs were willing to pay over $133,000 if the vehicle had 108 MPC. Current technology is unable to deliver this type of vehicle at any price. The results from our base case show that if price and MPC are in balance, there is a small niche of consumers who will be interested in electric vehicles.

In our first experiment, we found that CAFE is effective in diffusing alternative fuel vehicles. The overall sales of AFVs (both hybrid sedans and SUVs) increased significantly (more than 9%). Consumers bought more hybrids, but they bought more hybrid SUVs compared to the case when CAFE was not implemented. Thus, CAFE was ineffective in maximizing the social good (air pollution reduction) because as the fuel economy of SUVs improved they became more attractive to consumers. Additionally, CAFE had no impact on improving the fuel efficiency of gas sedan. Manufacturers design sedans at a 27.6 without CAFE, which is better fuel economy than the current CAFE regulation of 27.5. In addition, any penalties for SUVs that aren't in compliant to CAFE are passed on to the consumer who is willing to pay the price for their preferred vehicle. Thus, the effectiveness in CAFE of increasing the adoption of AFVs in order to improve the social good must be called into question.

In our second experiment, we looked at how technology push may impact consumers' purchasing decisions. We found that limiting the choice of types of engines to consumers did not

[2] June 12 2009, Ed Hellwig http://blogs.insideline.com/straightline/2009/06/gm-drops-mild-hybrid-sedans.html

cause them to move to the 'none' option. These results indicate that technology push can be important in the diffusion of AFVs, and with other eco-innovations.

We also considered the impact of word-of-mouth on the diffusion of AFVs in the third experiment. We found that WOM has a positive impact on the adoption of AFVs. Manufacturers of EVs found a small niche to satisfy and as the fuel efficiency of hybrid sedans improved compared to the base case, market share increased for these vehicles. WOM had a negative impact on the market share of both hybrid and gasoline SUVs. It should be noted that not all word-of-mouth is positive for each of the features. It is the combined product attributes (the product design) that delivers a vehicle that consumers are willing to embrace. For example, the WOM-adjustment factor is negative for hybrid engines, yet, in our simulation we saw an increase in market share compared to the base case. This is because the MPG improved from the base case, thus, offsetting the negative perception about hybrid engines. WOM can be a positive influence on speeding the diffusion of AFVs because consumers become more aware of the benefits of AFV ownership.

In this model we have several limitations. For example, we did not consider the relationship that manufacturers and consumers have with auto dealers. Although, there are always more complex models that could be developed using an ABM, Occam's razor is necessary in order to interpret the results. We take that approach with this initial model and look to add in greater complexity in future models. The insights that even simple ABMs can provide for understanding the mechanisms that drive co-dependent agents cannot be overlooked.

REFERENCES

Adner, R. and Levinthal, D. (2001), Demand Heterogeneity and Technology Evolution: Implications for Product and Process Innovation *Management Science* 47(5): 611-628.

Agrawal, A. W. and Dill, J. (2007). How to Pay for Transportation? A Survey of Public Preferences in California. Transport Policy 14(4): 346-356.

Andrews, R. L., Ainslie, A., and Currim, I. S. (2002). An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity. Journal of Marketing Research (JMR) 39(4): 479-487.

Bass, F. M. (1969). A New Product Growth Model for Consumer Durables. Management Science 15(January): 215-227.

Baumol, W. J. and Oates, W. E. (1988). The Theory of Environmental Policy Cambridge: Cambridge University Press

Beggs, S., Cardell, S., and Hausman, J. (1981). Assessing the Potential Demand for Electric Cars. Journal of Econometrics 16: 1–19.

Brownstone, D., Bunch, D. S., Golob, T. F., and Ren, W. (1996). A Transactions Choice Model for Forecasting Demand for Alternative-Fuel Vehicles, UC Davis: Institute of Transportation Studies (Ed.).

Brusco, M. J., Cradit, J. D., and Stahl, S. (2002). A Simulated Annealing Heuristic for a Bicriterion Partitioning Problem in Market Segmentation. Journal of Marketing Research 39(1): 99-109 (February).

Byrne, M. R. and Polonsky, M. J. (2001). Impediments to Consumer Adoption of Sustainable Transportation: Alternative Fuel Vehicles. International Journal of Operations & Production Management 21(12): 1521 - 1538.

Carley, K. M. (1996). Validating Computational Models, working paper Carnegie Mellon University.

Černý, V. (1985), Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm. Journal of Optimization Theory and Applications 45(1): 41-51 (January).

Chevalier, J. A. and Mayzlin, D. (2006). The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research (JMR) 43(3): 345-354.

Christensen, C., M (1997). The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Boston, MA: Harvard Business School Press.

Dellarocas, C., Zhang, X. M., and Awad, N. F. (2007). Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures. Journal of Interactive Marketing 21(4): 23-45.

Delucchi, M. (2005). AVCEM: Advanced-Vehicle Cost and Energy Use Model. Institute of Transportation Studies.

Duan, W., Gu, B., and Whinston, A. B. (2008). The Dynamics of Online Word-of-Mouth and Product Sales—An Empirical Investigation of the Movie Industry. Journal of Retailing 84(2): 233–242.

Garcia, R., Rummel, P., and Hauser, J. (2007). Validating Agent-Based Marketing Models through Conjoint Analysis. Journal of Business Research 60(8): 848-857.

Goldenberg, J., Libai, B., and Muller, E. (2002). Riding the Saddle: How Cross-Market Communications can Create a Major Slump in Sales. Journal of Marketing 66(April): 1-16.

---- (2001). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters 12(3): 211-223.

Golder, P. N. and Tellis, G. J. (2004). Growing, Growing, Gone: Cascades, Diffusion, and Turning Points in the Product Life Cycle Marketing Science 23(2): 207-218.

Herr, P. M., Kardes, F. R., and Kim, J. (1991). Effects of Word-of-Mouth and Product-Attribute Information of Persuasion: An Accessibility-Diagnosticity Perspective. Journal of Consumer Research 17(4): 454-462.

Kim, H. M., Rideout, D. G., Papalambros, P. Y., and Stein, J. L. (2003). Analytical Target Cascading in Automotive Vehicle Design. ASME Journal of Mechanical Design 125(3): 481-489.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). Optimization by Simulated Annealing. Science New Series 220(4598): 671–680.

Klemmer, P., Lehr, U., and Loebbe, K. (1999). Environmental Innovation, in Volume 3 of Publications from a Joint Project on Innovation Impacts of Environmental Policy

Instruments. Synthesis Report of a project commissioned by the German Ministry of Research and Technology (BMBF). Analytica-Verlag, Berlin.

Kopalle, P. K. and Lehmann, D. R. (1995),. The Effects of Advertised and Observed Quality on Expectations About New Product Quality. Journal of Marketing Research (JMR) 32(3): 280-290.

Lenk, P. J., Desarbo, W. S., Green, P. E., and Young, M. R. (1996). Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Design. Marketing Science 15(2): 173.

Loureiro, M. L., Mccluskey, J. J., and Mittelhammer, R. C. (2002). Will consumers pay a premium for eco-labeled apples? Journal of Consumer Affairs 36.

Meuter, M. L., Ostrom, A. L., Roundtree, R. I., and Bitner, M. J. (2000). Self-Service Technologies: Understanding Customer Satisfaction with Technology-Based Service Encounters. Journal of Marketing 64(3): 50-64.

Michalek, J. J., Feinberg, F. M., and Papalambros, P. Y. (2005). Linking Marketing and Engineering Product Design Decisions via Analytical Target Cascading. Journal of Product Innovation Management 22: 42-62.

Michalek, J. J., Papalambros, P. Y., and Skerlos, S. J. (2004). A Study of Fuel Efficiency and Emission Policy Impact on Optimal Vehicle Design Decisions. Transactions of the ASME 126: 1062-1070.

Midgley, D., Marks, R., and Kunchamwar, D. (2007). Building and Assurance of Agent-Based Models: An Example and Challenge to the Field. Journal of Business Research 60(8): 884-893.

Moorthy, K. S. (1985). Using Game Theory to Model Competition. Journal of Marketing Research 22(3): 262-282 (August).

Moreau, C. P., Lehmann, D. R., and Markman, A. B. (2001). Entrenched Knowledge Structures and Consumer Response to New Products. Journal of Marketing Research (JMR) 38(1): 14-29.

Rennings, K. (2000). Redefining Innovation — Eco-Innovation Research and the Contribution from Ecological Economics. Ecological Economics 32(2): 319-332.

Rossi, P. E., McCulloch, R. E., and Allenby, G. M. (1996). The Value of Purchase History Data in Target Marketing. Marketing Science 15(4): 321.

Sawtooth Software Inc. (2009), CBC Advanced Design Module of Sawtooth Software. Sequim, WA: Sawtooth Software, Inc. 530 West Fir Street, Sequim, WA 98382-3284 U.S.A., http://www.sawtoothsoftware.com/ (last update: January 26, 2010).

Toubia, O., Goldenberg, J., and Garcia, R. (2009). Diffusion Forecasts Using Social Interactions Data. MSI Reports Working Paper Series 09-210.

Urban, G. L., Weinberg, B. D., and Hauser, J. R. (1996). Premarket Forecasting of Really-New Products. Journal of Marketing 60: 47-60.

Weiss, A. M., Lurie, N. H., and MacInnis, D. J. (2008). Listening to Strangers: Whose Responses Are Valuable, How Valuable Are They, and Why? Journal of Marketing Research (JMR) 45(4): 425-436.

Whitehead, J. W. (2001). Design and Performance of Derivative-Free Optimization Algorithms Used with Hybrid Electric Vehicle Simulations. working paper University of Michigan.

Winebrake, J. J. and Farrell, A. (1997). The AFV Credit Program and its Role in AFV Market Development. Transportation Research Part D: Transport and Environment 2(2): 125-132.

THE IMPACT OF RESPONDENTS' PHYSICAL INTERACTION WITH THE PRODUCT ON ADAPTIVE CHOICE RESULTS

ROBERT J. GOODWIN
LIFETIME PRODUCTS, INC.

ABSTRACT

Lifetime Products, Inc., a manufacturer of folding furniture and other consumer hard goods, wanted to determine the potential impact of respondents' physical interaction with the product on the precision of adaptive choice (ACBC) results. Split-sample ACBC studies were conducted using online and mall-intercept field methods. Market simulation results were then validated using actual product sales and market share distributions. While online interviewing generally provided the most reasonable simulation share estimates, there could be cases (product novelty or complexity) where personal interviews with product interaction might be indicated.

INTRODUCTION

Lifetime Products, Inc. is a privately held, vertically integrated manufacturing company headquartered in Clearfield, Utah. The company manufactures consumer hard goods typically constructed of blow-molded polyethylene resin and powder-coated steel. Its products are sold primarily to consumers and small businesses worldwide through a wide range of discount department stores, home improvement centers, warehouse clubs, sporting goods stores, and other retail outlets.

Over the past four years, the Lifetime Marketing Research Department has adopted progressively more sophisticated conjoint and other quantitative marketing research tools to better inform product development and marketing decision making. The company's experiences in adopting and cost-effectively utilizing these sophisticated analytic methods – culminating in its current use of Sawtooth's Adaptive Choice-Based Conjoint (ACBC) software – were documented in a paper presented at a previous Sawtooth Software Conference. (Goodwin, 2009)

This paper first describes the research problem faced by Lifetime Products, that is, to determine whether the company's ongoing use of online surveying provides the most projectable conjoint results. Then the choice experiment is described, wherein the company conducted split-sample conjoint studies comprising both online and in-person (mall-intercept) field methods on its Folding Table & Chair category. Shares of preference from the conjoint simulation experiments were compared with actual sales distributions to determine which field method yielded results closest to actual.

I. THE RESEARCH PROBLEM

From a practical standpoint, there are key benefits from administering ACBC studies via the Internet. The online field method is well-suited to the adaptive nature of survey instruments, can utilize widely available consumer panels to generate representative samples, and allows for speedy completion of studies – all at competitive costs. Obviously, online respondents cannot "touch and feel" the physical product ("kick the tires," as it were), but for many well-known

product categories the lack of physical interaction is not a problem. Since in these instances the consumer already has a clear mental vision of the product, an onscreen rendering or photograph embedded in the questionnaire could suffice.

In Lifetime's case, though, such product understanding cannot always be assumed. Based on recent primary research results, the Company suspects that many of its products *do* need to be experienced physically – touch, sit in, set up, take down, move, etc. – before proceeding with an adaptive choice experiment. Many consumers find it difficult to project what a "blow-molded, high-density polyethylene and powder-coated steel" product really looks or feels like. And substitution of alternate (simplified) terminology, such as just "plastic and steel," might not help matters, since this could denigrate the image of the product unrealistically. Likewise, attributes such as perceived quality, ease of assembly or use, comfort, and subtle color or design aesthetics could be misinterpreted unless the product is experienced "hands-on."

Despite the benefits of realistic product appraisal inherent in mall-intercepts or other personal-interview formats, these field methods present a host of other concerns to the Company (e.g., more expensive, more difficult to obtain projectable samples, less control over the quality of respondents, longer turnaround times, logistics of shipping physical product to multiple venues, etc.). If it could be demonstrated that the online survey method yields ACBC model validation *equal to* (or better than) that of in-person methods, then the company would have the confidence to continue using the online method – with its relative speed, convenience, and cost savings – in future studies. (A summary of the key benefits and drawbacks of online and mall-intercept field methods for use with conjoint experiments is shown in Figure 1 below.)

Figure 1. Evaluation of Survey Options for ACBC Studies

	Online/Panel	In-Person/Mall-Intercept
Advantages	➢ Well-suited to adaptive instruments ➢ High-quality panel samples ➢ Speedy completion ➢ Competitive costs	➢ Touch-and-feel product ➢ Can use CAPI for ACBC
Disadvantages	➢ Have to use pictures & descriptions only	➢ Projectability issue ➢ Mall location selection ➢ Quality/consistency ➢ Ship product examples ➢ Takes longer to complete ➢ High costs

II. The Choice Experiment

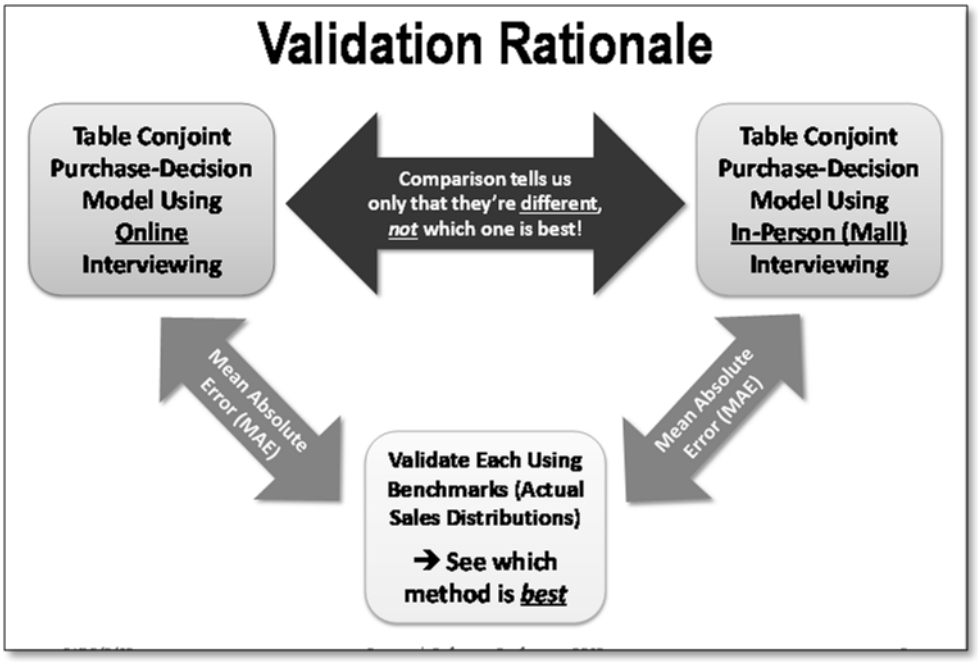
A search of previous research on the relative validity of online versus mall-intercept experimental studies yielded few examples. Jordan Lin conducted a similarly inconclusive literature search as part of his 2008 study of consumer response to food labels. However, in his study Lin found that online and mall-intercept results were comparable in main and interaction

effects. Further, he discovered that online data were relatively less vulnerable to social desirability bias, satisficing, and privacy concerns. (Lin, 2008)

Given the lack of breadth in the literature on this subject – but buoyed by Lin's findings – Lifetime decided to conduct split-sample adaptive choice research involving these two field methods. In late 2009 and early 2010, the Company conducted two such split-sample studies, one involving folding banquet/utility tables and the other with folding banquet/utility chairs. For both online and mall-intercept tracks, the product options were displayed using Sawtooth Software's ACBC conditional graphics feature. But in the mall-intercept track, respondents were also shown physical examples of the products with the invitation to touch, handle, and (in the case of the chair study) sit on the product before completing the CAPI-administered questionnaire.

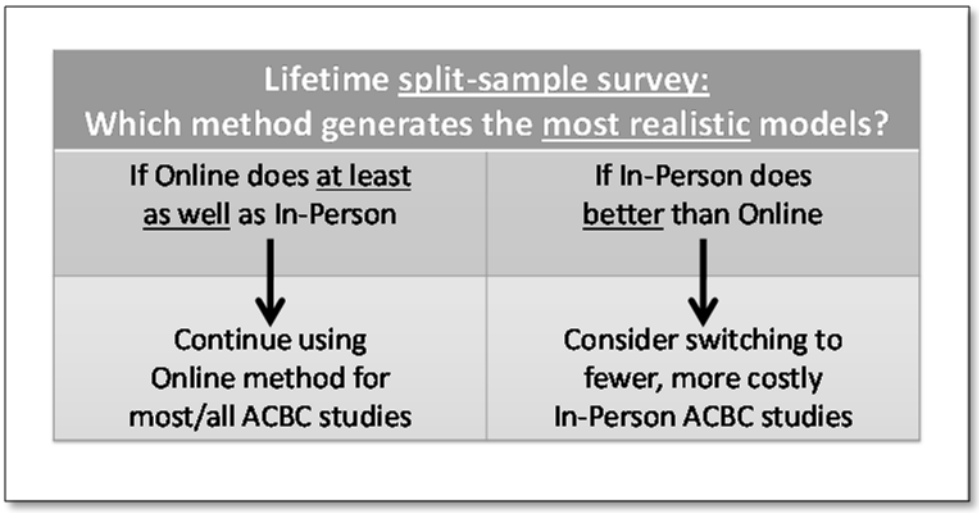
Validation Rationale. In order to validate the split-sample research, it was not enough merely to compare the results of each research wave with each other. We needed to benchmark the results against actual sales distributions to see which field method did a better job of "back-forecasting" actual sales (see Figure 2).

Figure 2



Because of the previously cited drawbacks in conducting mall-intercept surveys, we felt all we had to do was to demonstrate that the online method does *at least as well* as the in-person (mall-intercept) method. In essence, the burden of proof would be on in-person methods to demonstrate that they were superior (see Figure 3).

Figure 3. Burden of Proof in Split-Sample Experiment



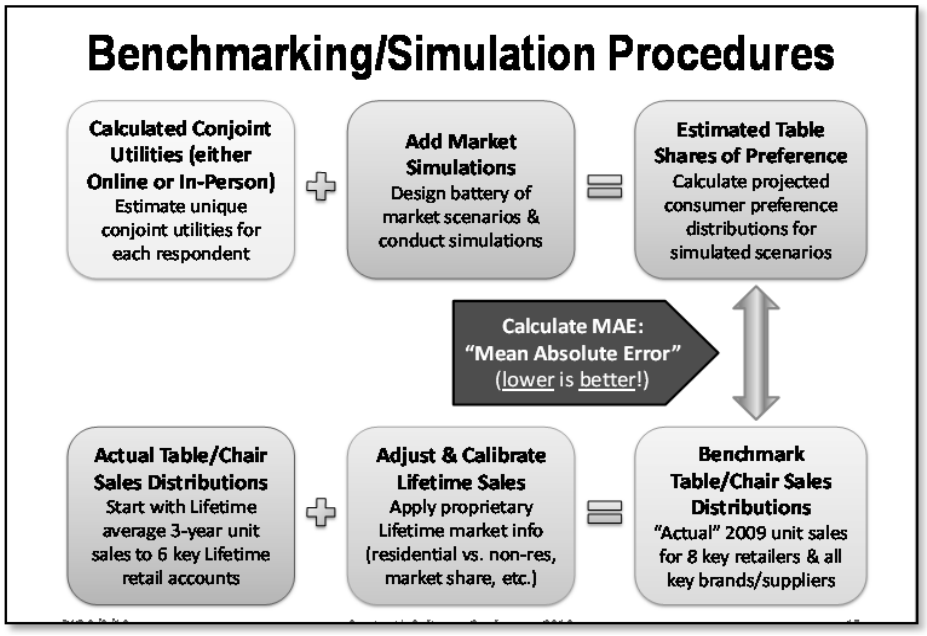
Benchmarking Procedure. Ideally, actual sales distributions would be available to benchmark market simulation results for each field method. However, in the case of the folding furniture category (and, indeed, most other Lifetime categories as well), actual sales distributions for the entire market are not available. The category is too small to warrant sales estimates by government or industry groups that are refined enough to be useful in the benchmarking process.

Fortunately, Lifetime has a relatively large share of market in the Folding Table & Chair category, so overall market shares can reasonably be computed using (a) current Lifetime sales distributions for the portions of the market it serves and (b) Company proprietary knowledge of portions of the market it doesn't serve. In using this method, the possibility for error still exists, but the magnitude thereof is probably not all that different from the challenges faced by many other industries who must supplement actual share data with market intelligence estimates.

The vast majority of retail folding furniture sales in the U.S. (about 85%) takes place in eight key retail chains: Three warehouse clubs (Sam's Club, Costco Wholesale, and BJ's Wholesale), three "big-box" discount department stores (Wal-Mart, Kmart, & Target), and two "DIY" home centers (Home Depot and Lowe's). Over the past three years, Lifetime has been a major supplier (in some cases the *sole* supplier) of resin-based folding furniture to six of these eight retailers. Therefore, Lifetime has considerable proprietary knowledge of the distribution of folding furniture sales in the market, with actual sales in six retail chains supplemented by reasonably sound market intelligence regarding the other two.

Because of this proprietary market knowledge, it was possible for us to construct "actual" unit sales distributions with acceptable precision. These sales distributions were then used as benchmarks for comparison with market simulation distributions driven by the respective conjoint models (see Figure 4).

Figure 4



Sampling Plans: The online tracks for these experimental studies used nationwide panel samples of approximately 400 subjects each. The in-person tracks were conducted in three more-or-less representative mall locations – Portland (Oregon), Chicago, and Atlanta – with total sample sizes of approximately 240 for each product. (Company research budget restraints prevented sampling larger numbers of mall-intercept consumers or using a wider array of mall locations.) Across both sets of products (tables and chairs) and field methods (online and mall-intercept), a total of 1,270 conjoint interviews were conducted for this split-sample test (see Figure 5).

Figure 5. Sample Characteristics (Total n=1,270)

Wave	Folding Banquet/Utility Tables	Folding Banquet/Utility Chairs
Online	• Nationwide panel sample • n=391 (±5.0% M/E @ α=.05) • February 2010	• Nationwide panel sample • n=397 (±4.9% M/E @ α=.05) • February 2010
In-Person	• Mall-intercept CAPI survey: • n=240 (±6.3% M/E @ α=.05): – Portland 80 (October 2009) – Chicago 80 (April 2010) – Atlanta 80 (April 2010)	• Mall-intercept CAPI survey: • n=242 (±6.3% M/E @ α=.05): – Portland 82 (October 2009) – Chicago 80 (April 2010) – Atlanta 80 (April 2010)

General respondent qualifications for these surveys were as follows:

- Male/Female (good mix)
- Age 25-64 (good mix)
- Homeowners (or stable renters for the mall-intercepts)

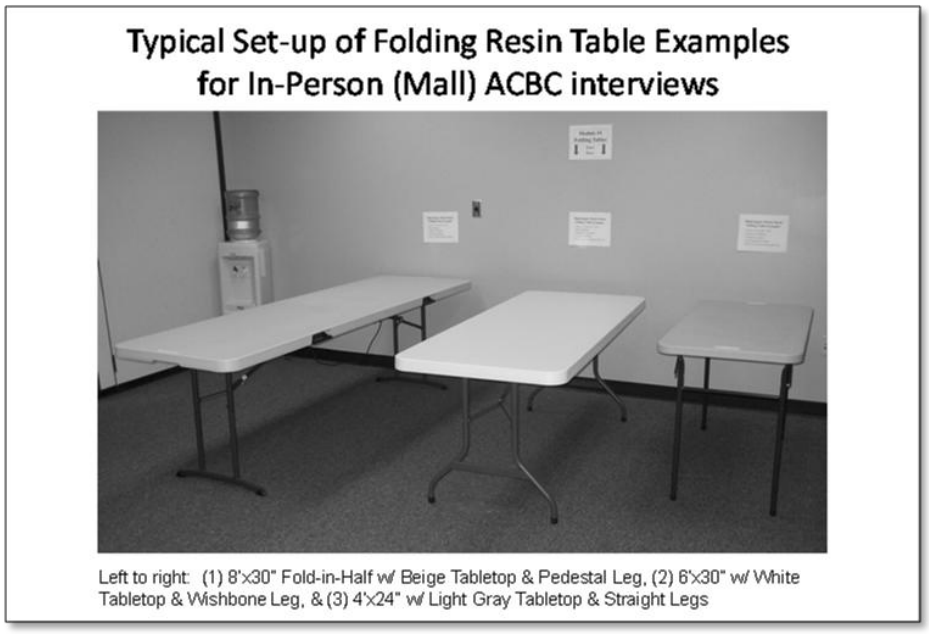
- Either *currently own* or are *planning to buy* in next 12 months one or more products in the respective test category (see visual examples shown to prospective mall-intercept respondents in Figures 6a & 6b)

Figure 6a

Figure 6b

Product Demonstration. During the mall-intercept phase of the folding table study, respondents were shown three table examples with varying size, leg style, color, and fold-in-half feature. Before proceeding with the Folding Table ACBC experiment, these respondents were asked to examine, touch, and lean on the table examples. All brand logos & other identifying marks were removed or concealed during examination (see Figure 7).

Figure 7



Likewise, those participating in the Folding Chair ACBC experiment in the malls were shown six representative chairs and asked to examine, touch, and sit in each one. The placement of these six chairs was randomized periodically to minimize potential order effects (see Figure 8).

Figure 8

Market Simulation Specifications: We used the Randomized First Choice (RFC) model in Sawtooth Software's SMRT package for all market simulations. We experimented with a range of exponential scale factors between 0.1 (flattened shares of preference) and 2.0 (accentuated shares of preference) in order to minimize the mean absolute error (MAE) for each simulation. We felt this scale factor adjustment process was warranted because of market aberrations in folding furniture retailing, in particular the widespread retail practice of limiting the number of brand/SKU offerings in each store. This practice tends to violate the assumptions of free market information and choice, and instead promotes consumer satisficing, product-substitution, and in some cases impulse-buying behaviors.

In order to account for non-buying tendencies by many consumers (i.e., they already have the tested products or they may not find their desired model configuration in the conjoint exercise), we used a "None" weight of 1. We then rescaled the resulting shares of preference for each product configuration to a total of 100% in order to derive shares of preference for those *most* likely to buy.

Finally, common retailing practices in this category (described later) encourage some interesting consumer purchase-decision behaviors, including satisficing, product substitution, and impulse buying. This prompted us to run a battery of market simulations using multiple market compositions in order to accommodate both consumers who buy whatever is available in the first store they shop *and* those who "shop around" for the best features and price.

III. Folding Table Experiment

Conjoint Design. The specification for the Folding Table ACBC experiment consisted of nine attributes with 31 levels, plus a price attribute with continuous (infinite) levels generated by

random experimental variation of ±25%. The complete design – including the respective price components for each level – is shown in Build-Your-Own (BYO) format in Figure 9.

Figure 9

Please design your ideal folding table by selecting one option for each feature.

(Note: Total price at bottom of page includes $29.99 base price.)

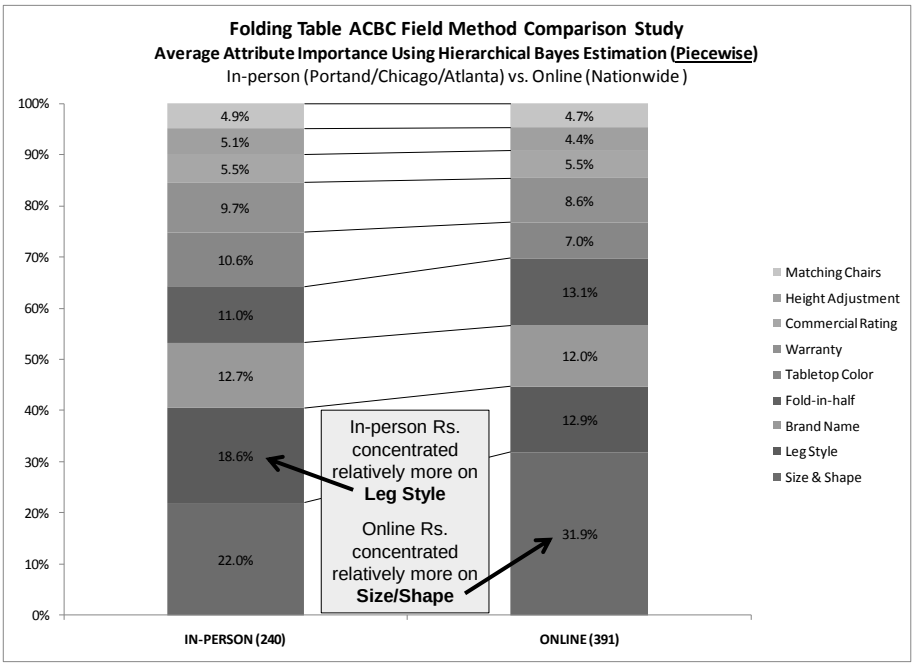
Feature	Select One Option for Each Feature	$ Add-on for Feature
Size & Shape	○ 4'x24" Rectangular ◉ 6'x30" Rectangular (+$10.00) ○ 8'x30" Rectangular (+$25.00) ○ 5-foot Round (+$50.00)	$ 10.00
Tabletop Color	○ White tabletop ◉ Beige tabletop ○ Light gray tabletop	$ 0.00
Leg Style	○ Straight legs ◉ "Wishbone" legs (+$1.00) ○ "Pedestal" legs (+$2.00)	$ 1.00
Fold-in-half	○ Does NOT fold in half ◉ Folds in half for storage (+$5.00)	$ 5.00
Height Adjustment	◉ Does NOT adjust (29" high only) ○ Adjusts to 22"/29"/36" high (+$5.00)	$ 0.00
Matching Chairs	○ Matching chairs NOT sold in same store ◉ Matching chairs sold in same store	$ 0.00
Commercial Rating	○ NOT commercial grade ◉ Commercial grade (+$3.00)	$ 3.00
Brand Name	○ NO brand name ○ Cosco brand (+$2.00) ○ Iceberg brand (+$2.00) ◉ Lifetime brand (+$2.00) ○ Meco brand (+$2.00) ○ Mity-Lite brand (+$8.00) ○ Office Star brand (+$2.00) ○ Rubbermaid brand (+$2.00) ○ Samsonite brand (+$2.00)	$ 2.00
Warranty	○ NO warranty ○ 5-year warranty (+$2.00) ◉ 10-year warranty (+$4.00) ○ Lifetime warranty (+$6.00)	$ 4.00
	TOTAL PRICE (including $29.99 base price)	$ 54.99

The heart of the conjoint model is the first attribute – "Size & Shape" – which includes the following array typical of many retail merchandising line-ups:

- The 6'x30" Rectangular Table is considered the "flagship" model in most folding table line-ups. As a key cross-over product, it is popular among both residential and commercial users.
- The 4'x24" Rectangular Table is also popular, especially among residential users. Its convenient size (especially with the fold-in-half option) gives it some impulse-buying characteristics.
- The 8'x30" Rectangular and (especially) 5-foot Round Tables are more commercial in nature, but are also seen in residential settings as well.

Attribute Importance and Utility Distributions. A comparison of the attribute importance distributions for each field method is shown in Figure 10. It should be noted that, although the two distributions are similar for each field method, the online subjects tended to put greater emphasis on Size & Shape of the table, while their mall-intercept counterparts paid relatively more attention to the table's Leg Shape.

Figure 10



The motivations for these differences in importance are shown in Figure 11. As shown in the Size & Shape call-out with the first graph segment, online respondents were much more discriminating than their mall-intercept counterparts in the selection of 6-foot and 8-foot Rectangular Tables and avoiding the 4-foot Rectangular and 5-foot Round models. In contrast, the Leg Shape call-out with the second graph segment, the mall-intercept respondents were relatively more likely to choose the "Wishbone" and "Pedestal" leg styles and to avoid the "Straight" leg style.

Figure 11

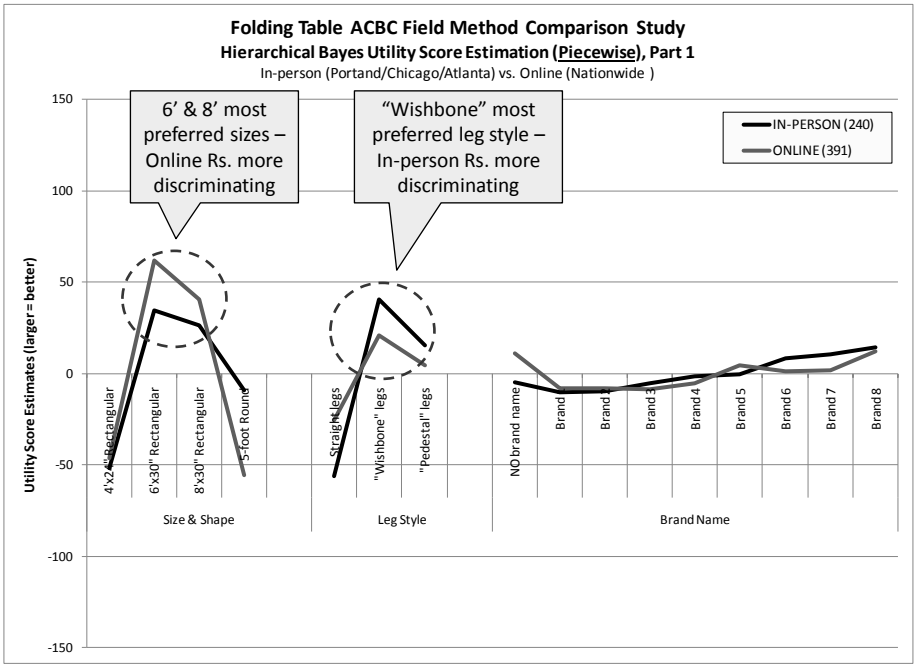


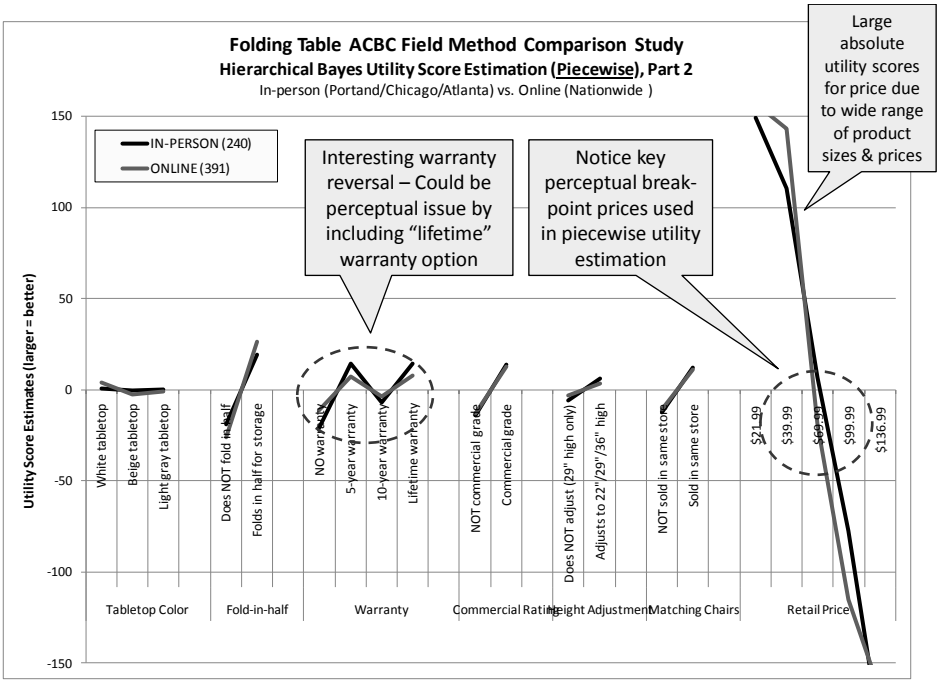
Figure 12 shows a part-worth graph for the other folding table attributes in this experiment. The most notable distribution is that of price in the call-out at far right. The extreme range of this attribute is due to the wide array of price options, from $21.99 for the lowest-price 4-foot Rectangular Table to $136.99 for the highest-price 5-foot Round Table (including the ±25% random price variation in the conjoint model). The piecewise price breaks were selected as follows:

$39.99 = approximate price break point for 4-foot vs. 6-foot rectangular tables

$69.99 = approximate price break point for 6-foot vs. 8-foot rectangular tables

$99.99 = approximate price break point for 8-foot rectangular vs. 5-foot round tables

Figure 12



Also of interest in this figure is the unexpected utility reversal in the Warranty attribute (see call-out with the third graph segment in Figure 12). This reversal may be due to the use of a combined alpha-numeric construct for the level descriptions (i.e., "5-year warranty … 10-year warranty … lifetime warranty"). We conjecture that when subjects viewed an array of product configurations in each task, they were confounded to some degree by the non-linear nature of these warranty descriptions and in so doing placed relatively less emphasis on the 10-year level. We have seen part-worth reversals similar to this in other studies involving a "lifetime" warranty option, as well as more consistent (expected) patterns where only numbers were used in the descriptions (see for example our Folding Chair exercise later in this paper).

Market Simulation Specifications. Due to (a) the common retail practice of product/SKU simplification and (b) the tendency of many consumers *not* to shop around (i.e., buy whatever is there), resulting in satisficing and substitution behaviors, we elected to conduct a variety of market simulations. We found that the five simulations summarized in Figure 13 represented a good overall view of the market.

Figure 13. Table Attributes Tested in Five Simulations

	Sim A Overall Market (11 configs.)	Sim B Simplified Market (7 configs.)	Sim C Typical 4-SKU (4 configs.)	Sim D 6' Rectangular (8 configs.)	Sim E 4' Rectangular (6 configs.)
1) Size/Shape	√	√	√		
2) Leg Style	√	√	√	√	
3) Brand				√	√
4) Fold-in-half	√	√		√	√
5) Color				√	√
6) Warranty				√	√
7) Rating	√	√	√	√	√
8) Height Adjustability	√	√		√	√
9) Matching Chairs				√	√
Retail Price	√	√	√	√	√

It will be noted that Simulations A to C excluded some attributes from the model. Exclusion of these attributes is not unlike what is commonly seen in many table line-ups, i.e., there is little if any variation in brand, color, warranty, or availability of matching chairs offered by a given retailer. In addition, some of these excluded attributes were less important in consumers' purchase decision making, so their exclusion resulted in a more parsimonious approach to the market simulation exercises.

Figure 13 also shows that Simulations D and E had the widest coverage of attributes, but focused only on the "flagship" 6-foot Rectangular and popular 4-foot Rectangular sizes, respectively. These two simulations were designed to capture potential non-compensatory purchase behaviors among consumers who want only a given table model and may shop around to get the configuration and price they strongly prefer.

The next two figures provide examples of specific table market simulations. Figure 14 shows Simulation A which includes the eleven most commonly found product configurations in the U.S. folding table market. (Keep in mind that, while this is a fairly comprehensive market simulation, it is also unrealistic, since only a few of these configurations are typically merchandised in any given store). The simulation includes variations in table size and shape, leg design, features, grade (residential or commercial emphasis), and price.

Figure 14. Simulation A – Overall Folding Table Market

Size/Shape	Features	Leg Style	Grade	Price	Share*
4’ Rectangular	None	Straight	Residential	$29.99	XXXXX
	Adjustable	Straight	Residential	$32.49	XXXXX
	Fold-in-Half	Straight	Residential	$32.49	XXXXX
	Adjustable FIH	Straight	Residential	$34.99	XXXXX
6’ Rectangular	None	Wishbone	Commercial	$48.99	XXXXX
	Adjustable	Wishbone	Commercial	$53.99	XXXXX
	Fold-in-Half	Wishbone	Commercial	$53.99	XXXXX
8’ Rectangular	None	Wishbone	Commercial	$73.99	XXXXX
	Fold-in-Half	Wishbone	Commercial	$78.99	XXXXX
5’ Round	None	Pedestal	Commercial	$99.99	XXXXX
	Fold-in-Half	Pedestal	Commercial	$104.99	XXXXX

* Expected shares of preference (based on actual/estimated residential market shares) were suppressed due to confidentiality.

Simulations B and C (not shown herein) were progressively more realistic simplifications of Simulation A. In Simulation B (Simplified Overall Market) the number of product configurations was reduced from eleven to seven by collapsing the designs with market shares less than 1% into the next closest designs with market shares of more than 1%. And, in Simulation C (Typical 4-SKU Offering) the configuration list was further reduced from seven to only four. This 4-SKU model (single configuration for each size) may be the most accurate reflection of merchandising reality, since retailer purchase agents typically buy no more than one SKU per table size in a given selling year.

Figure 15 shows Simulation D with an array of eight configurations of the “flagship” 6-foot Rectangular table. In this simulation, all attributes other than Size & Shape are varied to reflect the complete array of 6-foot tables offered among a variety of retailers. (Although not listed in the graphic, the other attributes – Leg Style, Tabletop Color, Height Adjustment, Availability of Matching Chairs, and Warranty – were set to be consistent with product offerings in each of the retail chains listed.) Simulation E (not shown herein) was designed in like manner for six configurations of the popular 4-foot Rectangular table.

Figure 15. Simulation D – 6-foot Rectangular Table Detail

Retailer	Brand	Grade	Features	Price	Share*
Sam's Club	Lifetime	Commercial	None	$48.99	XXXXX
Costco Whls.	Lifetime	Commercial	Fold-in-half	$52.99	XXXXX
BJ's Wholesale	Private Label	Commercial	Fold-in-half	$52.99	XXXXX
Wal-Mart	Private Label	Residential	Fold-in-half	$44.00	XXXXX
Kmart	Private Label	Residential	None	$44.99	XXXXX
Target	Private Label	Residential	Fold-in-half	$49.99	XXXXX
Home Depot	Private Label	Residential	None	$49.98	XXXXX
Lowe's	Samsonite	Residential	None	$49.98	XXXXX

* Expected shares of preference (based on actual/estimated residential market shares) were suppressed due to confidentiality.

Simulation Validation. The results of each of the five Folding Table simulations (A-E) conducted using both online and mall-intercept ACBC models were validated by comparing shares of preference with actual/estimated market shares. Two diagnostics were employed, Mean Absolute Error (MAE) and Coefficient of Determination (R^2). A summary of the comparison of these key indicators is included in Figure 16.

Figure 16. Folding Table Simulation Validation Summary

Simulation Scenario (Scale Factor Exponent)	MAE Online	MAE In-Person	MAE Advantage for Online	R^2 Online	R^2 In-Person	R^2 Advantage for Online
A) Overall Market (Scale Factor = 1.5)	6.6%	7.4%	**-0.8%**	.484	.553	**-.069**
B) Simplified Market (Scale Factor = 1.5)	8.5%	10.0%	**-1.5%**	.256	.299	**-.043**
C) Typical 4-SKU Line-up (Scale Factor = 1)	7.1%	6.1%	**+1.0%**	.930	.941	**-.011**
D) 6' Rectangular Detail (Scale Factor = 0.25)	3.7%	5.9%	**-2.2%**	.437	.151	**+.286**
E) 4' Rectangular Detail (Scale Factor = 0.25)	2.4%	2.8%	**-0.4%**	.776	.704	**+.072**
Average of All Simulations	**5.6%**	**6.4%**	**-0.8%**	**.577**	**.530**	**+.047**
Weighted Average* of All Simulations	**5.7%**	**6.7%**	**-1.0%**	**.802**	**.806**	**-.004**

* MAEs weighted by number of configurations in each simulation scenario; R-squares based on combined correlation of all 36 configurations across all five simulations.

Looking first at the MAEs, the online model had lower error rates in four of the five simulations (all but C – Typical 4-SKU Line-up, where the mall-intercept model had less error).

The largest difference in favor of the online model (-2.2%) was in Simulation D – 6' Rectangular Detail dealing with the "flagship" table product. Using a simple arithmetic average of the MAEs across all five simulations, there was a slight performance advantage (-0.8%) for the online over mall-intercept models. Weighting the MAEs by the number of configurations in each scenario (11 in A, 7 in B, etc.) strengthened the online advantage slightly to -1.0%.

The individual levels of these MAEs (using optimum scale factors) ranged from 2.4% to 10.0%, with an arithmetic average of 6.0%. On face value this was a bit higher than the 3-4% range we would have expected (or liked). Some of this larger-than-expected error may be attributed to the fluidity of retailers' product line-up from year to year and the use of "estimated actual" market shares as benchmarks. However, the more likely cause is probably retailers' limited-SKU merchandising practices (discussed previously) were in scenarios with the most unrealistic merchandising arrays (Sims A and B).

Analysis of the direction of the errors in Simulations A and B reveals *under*estimation of standard, non-feature-rich product configurations (that are *more* likely to be offered by retailers) and *over*estimation of more desirable configurations including features as fold-in-half or height-adjustment (that are *less* likely to be offered). In like manner, the direction of errors for Simulation D shows *under*estimation of lower-priced, residential-grade 6-foot tables (which are typically sold through wide-distribution discount store networks) and *over*estimation of higher-priced, commercial-grade configurations (which are typically sold through limited-distribution warehouse club networks). These findings suggest that overall consumer satisfaction might be enhanced if (a) retailers added additional, high-value SKUs and (b) these products were more available in non-warehouse-club channels.

As suggested in feedback to the presentation of this paper in the conference, coefficients of determination (R^2s) are included in this analysis to provide a different perspective on the validation metrics. As shown in Figure 16, the arithmetic average of R^2 for the five simulations is .577 for the online model and .530 for the mall-intercept model, a slight advantage (+.047) for online. Combining all 36 configurations across the five simulations, the R^2 values jump to the low 80s, with a nominal (+.004) advantage for the mall-intercept model. Because of the differences in the methodologies of these diagnostic measures, the individual simulation-wise R^2 comparisons do not always correlate well with the corresponding MAEs (although the largest MAE advantage for online – Simulation D – did have the strongest R^2 advantage).

IV. Folding Chair Experiment

Conjoint Design. The specification for the Folding Chair ACBC experiment consisted of six attributes with 26 levels, plus a price attribute with continuous (infinite) levels generated by random experimental variation of ±25%. The complete design – including the respective price components for each level – is shown in Build-Your-Own (BYO) format in Figure 17.

Figure 17

Please design your ideal folding chair by selecting one option for each feature.

(Note: Total price at bottom of page includes $11.99 base price.)

Feature	Select One Option for Each Feature	$ Add-on for Feature
Chair Style	○ All Metal ○ Single-wall Plastic (+$2.00) ○ Padded Fabric (+$3.00) ◉ Contemporary Double-wall Plastic (+$4.00) ○ Classic Double-wall Plastic (+$4.00) ○ Mesh (+$8.00)	$ 4.00
Color	○ White ◉ Beige ○ Light gray ○ Black	$ 0.00
Matching Tables	○ Matching tables NOT sold in same store ◉ Matching tables sold in same store	$ 0.00
Commercial Rating	○ NOT commercial grade ◉ Commercial grade (+$1.50)	$ 1.50
Brand Name	○ NO brand name ○ Cosco brand (+$1.00) ◉ Lifetime brand (+$1.00) ○ Meco brand (+$1.00) ○ Mity-Lite brand (+$4.00) ○ Office Star brand (+$1.00) ○ Rubbermaid brand (+$1.00) ○ Samsonite brand (+$1.00)	$ 1.00
Warranty	○ NO warranty ○ 1-year warranty (+$1.00) ○ 5-year warranty (+$1.50) ◉ 10-year warranty (+$2.00)	$ 2.00
	TOTAL PRICE (including $11.99 base price)	$ 20.49

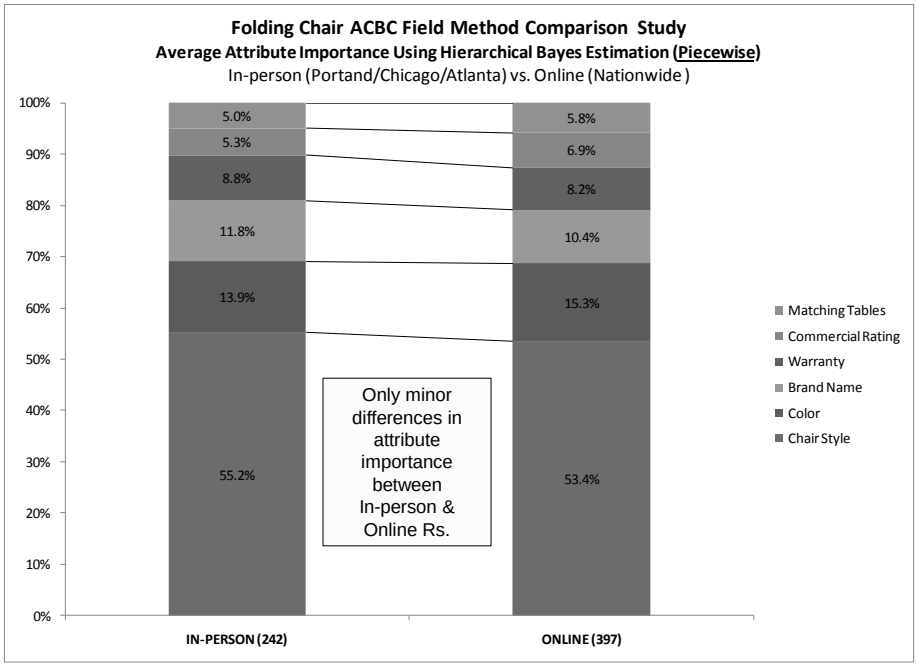
The heart of the conjoint model is the Chair Style attribute, which includes six key chair types available in most retailer line-ups:

- The All-Metal & Single-Wall Plastic chairs are low-cost, generally less-comfortable offerings.
- The Padded Fabric chair is a popular choice among consumers, though it does not match most folding tables sold by the same retailer.

- The two Double-Wall Plastic chairs represent a key segment in which Lifetime is a significant player. These chairs are ergonomically designed (more comfortable than they may look), typically commercial-grade, and match the folding resin tables generally sold at the same retailer. The Classic design has been on the market for more than a decade, while the Contemporary design is a newer, somewhat sleeker variant.
- The Mesh chair is new entrant in the market by one of Lifetime's competitors. Although it is potentially a strong player in the market (hence the reason it was included in this conjoint model), it has virtually no sales history, which creates some problems in the validation process (to be discussed later).

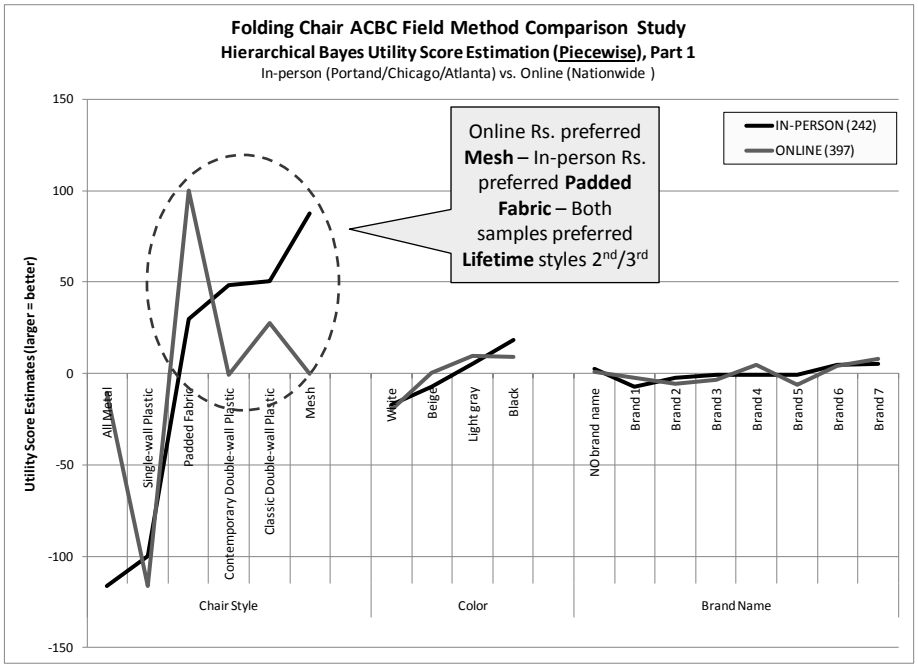
Attribute Importance and Utility Distributions. A comparison of the attribute importance distributions (excluding price) for each field method is shown in Figure 18. In contrast to the previous analysis on Tables, the two distributions for the Chair conjoint have only minor differences. In both cases, however, Chair Style was by far the most important attribute to consumers.

Figure 18



The graphic call-out on the left in Figure 19 shows that, while respondents in both field waves attached equally high importance to the attribute Chair Style, they did it for markedly different reasons. Subjects in the online wave preferred the Padded Fabric chair by a wide margin over all other styles. In contrast, mall-intercept respondents were so impressed with the comfort of the Mesh chair during the pre-conjoint comfort test that they ranked it first among the six options. Also note that, in both tests, the Double-Wall Plastic chair models were ranked second and third, with mall-intercept respondents tending to like it better than their online counterparts.

Figure 19



These results are consistent with the general direction of other Lifetime chair research studies, which found the following:

- Consumers like the idea of a padded fabric chair, but do not necessarily like the actual product when they try it out.
- In contrast, many consumers are not initially impressed with the concept of a (hard) Double-Wall Plastic chair, but are pleasantly surprised with the comfortable, ergonomic design of the Lifetime models.
- The Mesh chair, while high desirable to those who tried it out in the mall-intercept wave, did not sound impressive to those in the online wave, perhaps conjuring visions of outdoor patio furniture with less substantial mesh or weave construction.

Figure 20 shows part-worth graphics for other folding chair attributes in this experiment. Note that the range of the price attribute is not as extreme as in the Table experiment. The piecewise price breaks were selected as follows:

$14.99 = approximately the 1st Quartile for folding chair prices

$19.99 = a key perceptual price barrier for double-wall plastic folding chairs

$24.99 = approximately the 3rd Quartile for folding chair prices

Figure 20

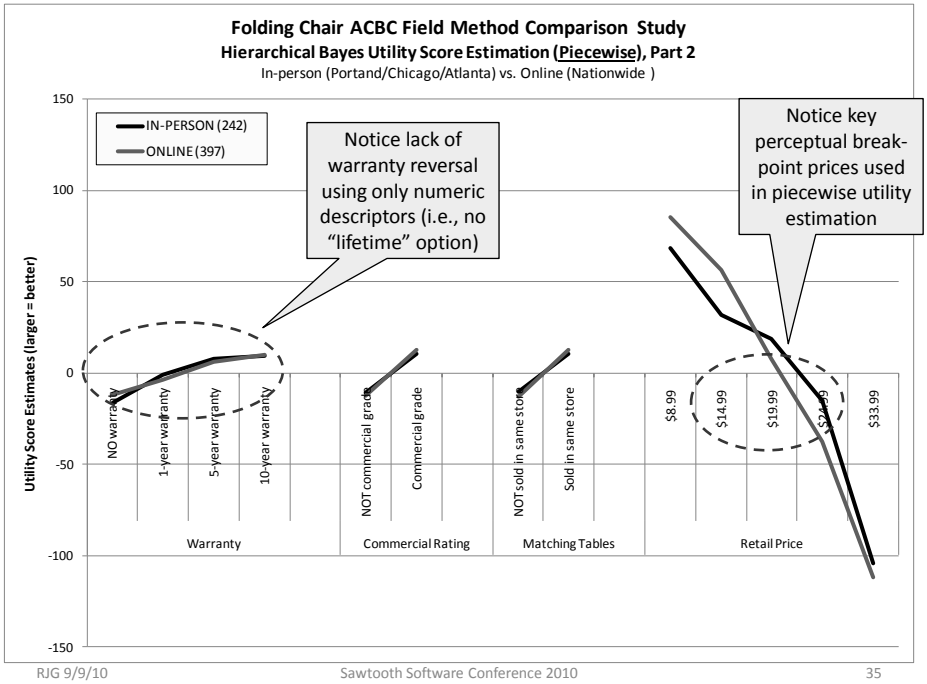


Figure 20 also shows that, in contrast to the utility reversal in the Table conjoint, the Warranty part-worths for chairs are consistently rising (see first graph segment). As discussed previously, this may be due to the use of a completely numeric set of level descriptions ("1-year warranty … 5-year warranty … 10-year warranty") which avoids adding the non-numeric "lifetime" description.

Market Simulation Specifications. It will be recalled that the merchandising practices in the folding table market (single-SKU offerings) necessitated the inclusion of multiple simulation scenarios to test for differences between online and mall-intercept field methods. Fortunately, in the folding chair category, retailers are much more likely to offer multiple design options in the same in-store line-up (sometimes up to four or five of the six designs tested in the study). Consequently, the case for numerous chair simulation scenarios to account for market "wrinkles" was not as compelling, and it was anticipated that perhaps a single overall-market simulation (see Simulation A in Figure 21) would suffice for this exercise.

Figure 21. Simulation A – Overall Folding Chair Market

Style	Color	Brand	Warranty	Retailer(s)	Price	Share*
Contemp.	Beige	Lifetime	10 years	Costco Whlse.	$19.99	XXXXX
Classic	White	Lifetime	10 years	Sam's Club	$19.64	XXXXX
	Beige	Pvt. Label	10 years	BJ's Whlse.	$19.99	XXXXX
	Beige	Samsonite	5 years	Home Cntrs.	$19.97	XXXXX
Pad. Fabric	Beige	Cosco	5 years	Sam's Club	$17.88	XXXXX
	Beige	Samsonite	5 years	Costco-BJ-Lowe's	$17.99	XXXXX
Pad. Vinyl	Black	Cosco	None**	Discounters (3)	$16.50	XXXXX
Metal	Beige	Cosco	1 year	Sam's Club	$11.69	XXXXX
	Gray	Cosco	1 year	WM-Home Cntrs.	$9.50	XXXXX
	Black	Cosco	1 year	Target	$9.59	XXXXX
Plastic	Black	Cosco	1 year	Discounters (3)	$9.00	XXXXX

* Expected shares of preference (based on actual/estimated residential market shares) were suppressed due to confidentiality.

** Padded Vinyl configuration was not included in the ACBC design, so for the simulation "No Warranty" was used with the Padded Fabric design to compensate. Post-simulation error analysis verified that this individual test configuration did not have a large impact on MAEs for either field method.

Nevertheless, the findings of this simulation were so unexpected (see analysis to follow) that we elected to add two sequential simplifications of this market structure (Simulations B and C) to re-test the direction and magnitude of the differences. The list of products was simplified from 11 configurations in Sim A to eight in Sim B (collapsing colors and brands) and further to six in Sim C (collapsing to a single configuration per chair style). Matching Tables were assumed to be available for the Classic and Contemporary Double-Wall Plastic styles, but not for any of the other configurations. Commercial Grade was assumed for each of the three Classic and Contemporary chair configurations offered by the warehouse club stores; Residential Grade was assumed for each of the remaining non-club configurations.

Of particular note is the absence of the Mesh chair from these simulations. This model was recently introduced by one of Lifetime's competitors and has had (as of this writing in November 2010) no obvious visibility in any large of the large retail chains (and therefore presumably no substantive sales history). Exclusion of this model from the simulations will have (as we shall see next) a sizable impact on the validation process for these conjoint experiments.

Chair Simulation Validation. As with the Folding Table phase, the results of each of the three Folding Chair simulations (A through C) using both online and mall-intercept ACBC models were validated by comparing shares of preference with actual/estimated market shares. A summary of the comparison of these key indicators is included in Figure 22.

Figure 22. Folding Chair Simulation Summary

Simulation Scenario (Scale Factor Exponent)	MAE Online	MAE In-Person	MAE Advantage for Online	R^2 Online	R^2 In-Person	R^2 Advantage for Online
A) Overall 11-SKU Array (Scale Factor = 1 for Online; 0.1 for In-Person)	3.6%	5.9%	**-2.3%**	.530	.032	**+.499**
B) Simplified 8-SKU Array (Scale Factor = 1 for both methods)	1.8%	9.0%	**-7.1%**	.911	.170	**+.741**
C) Typical 6-SKU Array (Scale Factor = 0.5 for Online; 0.1 for In-Person)	2.8%	7.8%	**-5.0%**	.924	.154	**+.770**
Average of All Simulations	**2.7%**	**7.6%**	**-4.9%**	**.788**	**.119**	**+.670**
Weighted Average* of All Simulations	**2.8%**	**7.3%**	**-4.9%**	**.882**	**.071**	**+.810**

* MAEs weighted by number of configurations in each simulation scenario; R-squares based on combined correlation of all 25 configurations across all three simulations.

As shown in this tabular display, the online conjoint design does a fairly good job of projecting to actual market shares for the Folding Chair category. The MAEs are all respectable, starting with 3.6% for the overall market view (Sim A) and improving somewhat for more simplified (and more realistic) market views (Sims B and C). The R^2s are especially strong in the simplified market structures, with the conjoint model shares of preference explaining about 90% of the actual shares of market.

The most striking feature of this validation analysis is the extremely poor performance of the mall-intercept conjoint model, in terms of both MAEs and (especially) R^2s. A priori, we expected that the folding chair experiment would be an opportunity for the mall-intercept wave to "shine," since actual experience with the product should be more compelling with chairs than with tables. But the MAEs were fairly large and R^2s were extremely small, indicating a very poor correlation between actual and projected shares. In fact, the apparent improvement in mall-intercept R^2s in Sims B and C (from .032 to the mid-teens) is deceptive, because their constituent Rs (Pearson's Correlation Coefficients) were unexpectedly *negative* for both of these simulations, indicating an *inverse* relationship between projected and actual shares.

Clearly, there is something else going on here – and it is almost certainly the absence of the Mesh chair from the validation analysis. The Mesh chair was by the far the favorite chair design for the mall-intercept participants, all of whom had a chance to try out all six chair styles before completing the conjoint experiment. However, when it came time to simulate their purchase behavior, their favorite chair design was not available, so (depending on the cross-elasticities among their part-worths) they were either allocated to one of the other chair styles for simulated

purchase or (more likely) they reverted to the None option, in effect saying none of the remaining chair styles were of interest for them to purchase. In fact, the None share of preference for the mall-intercept Chair experiment was quite a bit larger than the None shares for each of the other ACBC experiments in this study, placing much greater volatility in this particular validation analysis. In reality, the absence of the Mesh chair from the simulation array virtually guaranteed that that the mall-intercept method could not "beat" the online method in the chair study.

Discussion of Key Validation Issues

"Touch-and-feel": Less important than previously thought? The conventional wisdom is that respondents' physical interaction with table and (especially) chair products should improve the accuracy of conjoint utility estimation. In these two studies, however, the benefit of seeing and touching physical folding table examples apparently did not overcome the negative impacts of other factors (see below). It may indeed be that – for these categories, at least – consumers have enough savvy to be able to deal realistically with purchase decisions without having physical product examples nearby.

Differential quality of respondents. Online panels have increased in quality and "representative-ness" over the past decade. And, online purchasers tend to be more savvy and discriminating than the average consumer, so they have the ability to make rational purchase decisions without necessarily seeing and/or touching the physical product. In contrast, the quality of mall-intercept surveys (realistically the most cost-effective form of in-person quantitative surveying) remains questionable.

Potential for respondent fatigue during mall-intercepts. In order to hold mall field costs down, we jointly recruited for both table and chair surveys. Since most respondents owned both folding tables and folding chairs, about three-fourths of the mall-intercept sample completed *both* surveys, spending a total of about 30 minutes to do so. Although the order of presentation was rotated to minimize order bias, there was still the possibility of respondent fatigue. **T**he mall respondents had nearly the same survey-taking experience (other than seeing "live" product examples) by completing a self-administered computer-assisted personal interview (CAPI), which presented a nearly identical survey-taking experience (including the use of conditional graphics to enhance the realism of task choices) as that of the online respondents.

The nature of folding table purchase decision-making. Since folding tables are relatively non-complex, mature hard goods, consumers may be able to arrive at a purchase decision without the need for physical stimuli. For instance, upon seeing a print advertisement for a 6-foot folding table, a consumer may need to look for only a few cues or "signals" (e.g., "durable plastic resin," "folds in half," "10-year warranty," "$XXX price," etc.) in order to decide to purchase such an item the next time he/she visits the retailer. Most of the purchase decision may have been made before actually seeing the product in the store.

Inclusion of a new-product configuration in the Chair conjoint model. A new Mesh folding chair (recently introduced by one of Lifetime's competitors) was included in the conjoint model. It was highly regarded by the mall-intercept respondents (ranked first by a wide margin because of its comfort), but ranked only in the middle of the pack by the online respondents (who apparently had comfort and/or quality concerns with the simple "mesh" description and who instead opted for the "padded fabric" chair as their number-one pick). Because there is virtually

no sales history for the Mesh chair configuration, benchmarking was impossible and the design was therefore excluded from all of the simulations. Had there been actual sales history for the Mesh chair (making it possible to include in the simulations and compare it against accurate benchmarks), it is probable that the MAEs and R^2s for the mall-intercept wave would have improved markedly. Or, looking at it a different way, if the Mesh chair had been excluded *entirely* from the conjoint model (and not seen or tested by any of the respondents), it is also possible that the chair style utilities (particularly for the mall-intercept wave) would have been quite different, again with possibly ameliorating effects on the validation.

Conclusions

What Lifetime Products has learned from this research study. We continue to like the online field method, at least for Lifetime's standard, fairly familiar product lines such as folding banquet/utility tables and chairs. Online (panel) respondents are savvy enough to understand the product without the need to see the products. And, we avoid the higher costs of mall-intercept interviews (which in this study were *three times* as high on a cost-per-interview basis as the online method).

The Mesh chair "outlier" experience gives us pause. The lack of validation with sales benchmarks probably kept the mall-intercept method from "winning" the Chair study. The "Mesh Chair" description was probably insufficient to allow online participants to evaluate the chair design on the same plane as their mall-intercept counterparts. Generalizing this finding, we would conclude that new, innovative, or unfamiliar products may indeed need to be demonstrated in-person, despite the potential cost penalties and logistical hurdles of doing so.

In general, benchmarking process worked OK for us, but... "The Market" was somewhat illusive. Hard market data are not generally available in the folding table and chair segments in which Lifetime competes. Therefore, we had to make a number of assumptions in order to come up with reasonably accurate estimations to benchmark against conjoint simulation results. At the same time, retailer merchandising practices such as limited SKU offerings and annual model line-up changes presented several other analytic challenges. Because of these market "wrinkles," we elected to employ multiple conjoint simulations, both to explore different market "angles" separately and to aggregate them into an overall market view. Although the market uncertainties we experienced are probably not all that different from those of many other industries, perhaps there are some industries where more precise, well-known market share benchmarks can be used for this type of conjoint validation.

Where does Lifetime go from here? Lifetime may conduct additional conjoint validation test(s) with other categories using this split-sample approach. However, we'd do some things differently the next time around.

- **Avoid conjoint designs and product configurations that can't be benchmarked properly**. The Mesh chair was essentially "new to the world." It was innovative enough that it probably could not be evaluated realistically without interacting physically with the product. Because of this, the concept was assessed quite differently by online and mall-intercept respondents, creating markedly different conjoint models. And, the chair had no actual sales history, so it could not be validated properly. (As noted before, it was strategically important for Lifetime to evaluate consumer appeal of the new Mesh chair, hence its inclusion in this study.)

- **Avoid testing product "outliers" with low incidence of usage**. The 5-foot Round Table is targeted primarily to commercial users and has only limited appeal among consumers. Use of MAEs for validation in this case may be deceiving, since small MAEs may results in large MAPEs (Mean Absolute Percentage Errors). In addition, the higher average price of this table size extended the overall price range for the category, creating at least the potential for price cross-elasticity dislocations in the Table study.
- **Refine our estimation process for benchmarking**. We tried a number of approaches to make our actual sales distribution estimates as realistic as possible. However, since this could be a potential "Achilles' heel" in the validation process, further refinement efforts in future studies could be beneficial.
- **Include more product examples to touch-and-feel in the mall-intercept wave**. In the current study, three table examples and six chair examples were provided for mall respondent interaction. Much has been said above regarding the chairs, but a comment regarding the table study would be appropriate here. The three tables were selected to provide a variety of attribute levels (size, tabletop color, leg style, and fold-up/height-adjustment features). Obviously, the assumption was the consumers could mentally interconnect these features so that, say, the fold-up feature could be visualized on any size table or with any leg style. This mental interconnection probably could be made easier by (as interview space permits) showing a greater variety of product options.
- **Increase the sample size and geographical coverage in the mall-intercept wave**. In each of the current studies we surveyed 400 subjects online and only 240 total in three (fairly representative) malls. In future studies, it may be well to increase the mall wave to 400 total interviews (say, 80 in each of *five* malls) to provide a degree of statistical accuracy comparable to that of the online wave. (Obviously, the research budget for the mall wave would need to go up by about two-thirds to accomplish this enhancement.)
- **Use pre-recruited in-person interviews rather than mall-intercepts**. In our view, one of the key drawbacks to the use of mall-intercept personal interviews is the overall quality of respondents. As research budgets permit, Lifetime will consider pre-recruiting subjects for at least part of the In-Person wave in future validation studies (without having to reduce samples sizes unrealistically). Or, as suggested by Chris Chapman in his discussant remarks regarding the presentation during the conference, we might consider piggybacking conjoint experiments on focus group interviews.

REFERENCES

Goodwin, Robert J.: Introduction of Quantitative Marketing Research Solutions in a Traditional Manufacturing Company: Practical Experiences. 2009. *Proceedings of the Sawtooth Software Conference, pp. 185-198*.

Lin, Chung-Tung J.: Mall-Intercept vs. Online Panel – Does Sample Source for an Experimental Study Matter? 2008. *Paper presented at 63rd AAPOR Annual Conference*.

Orme, Bryan & Johnson, Richard: External Effect Adjustments in Conjoint Analysis. 2006. *Sawtooth Software Research Paper Series*.

USING EYE TRACKING AND MOUSELAB TO EXAMINE HOW RESPONDENTS PROCESS INFORMATION IN CBC

MARTIN MEIßNER,
SÖREN W. SCHOLZ,
AND REINHOLD DECKER
BIELEFELD UNIVERSITY, GERMANY

ABSTRACT

Respondents' attention and information processing are most often invisible for market researchers conducting conjoint studies. Cognitive psychology has shown that most respondents limit information processing when making choices. All the more, this kind of data is interesting to understand how respondents come to their final decisions. This paper investigates respondents' information acquisition behavior by means of process tracing techniques. The contribution of the paper is threefold: First, it discusses how respondents' attention is related to final choices. Second, it investigates whether and how attentional data can improve the validity of choice models. Third, it seeks to answer the question which process tracing approaches can adequately be used in conjunction with Choice-based Conjoint Analysis.

INTRODUCTION

Choice models are marketers' favorite for quantifying the influence of product attributes on consumer decisions. One of the most widespread approaches is Choice-based Conjoint Analysis (CBC). CBC statistically relates product attributes and levels to respondents' decisions, but neglects the cognitive processes taking place during the evaluation of choice tasks. This means that in most applications only respondents' final decisions will be used to calculate part-worth utilities and subsequently build models for market share prediction. Information processing and information integration are usually not investigated. However, from a behavioral perspective a multitude of cognitive processes take place before respondents come to their final decisions. These process-steps include perception (i.e. visual attention), cognition and behavioral selection (Logan and Zbrodoff 1999). Due to the fact that attention is a pre-conscious process to the final decision, it should be investigated in more detail to understand its relation to choice.

Psychological research has suggested sophisticated process tracing techniques, e.g. eye tracking (Lohse and Johnson 1996) and Mouselab (Jasper and Shapiro 2002), to better understand how decision makers acquire and integrate relevant information (see also Payne, Bettman, and Johnson 1993). Eye tracking is used to record respondents' eye movements and allows "fine-grained measurement of natural attentional flow and intensity" (Pieters and Warlop 1999 p. 13). In this way the experimenter is able to see the choice tasks with the eyes of the respondents. Analogously, Mouselab provides processing data by recording mouse movements.

In his wish list for conjoint analysis, Bradlow (2005) claimed that a better understanding of the processes taking place in the mind of the respondent is important for the development of better conjoint models. Other researchers noted that process data like eye and mouse movements, click-stream data and brain images might be utilized in preferences measurement (Netzer et al.

2008). Following these propositions the aim of this paper is to comprehensively examine the information acquisition processes in choice tasks because this information might help to develop more valid choice models. We therefore investigate the attentional processes in a typical CBC setting.

From the preference measurement perspective three issues are of major interest: First, we are interested in how the attentional processes are connected to the final choices. Previous studies have shown that the selection of information already includes preference information (Glaholt and Reingold 2009; Pieters and Warlop 1999; Shimojo et al. 2003). Therefore, we compare the part-worth utilities from CBC with the amount of information acquisition for attributes and levels. The second issue is whether data from the attentional process can be used to improve choice models. For this purpose eye tracking data are incorporated into the standard Hierarchical Bayes MultiNomial Logit (HB-MNL) model. Finally, we compare the data of both eye-tracking and Mouselab process tracing in order to answer the question whether Mouselab is suitable for applications in marketing research practice.

The remainder of the paper is structured as follows: Next, we give a brief overview on empirical studies investigating the relationship between attention and choice. Then, we outline the main principles of eye tracking and Mouselab and discuss advantages and disadvantages of both process tracing techniques. Following the presentation of the design and results of our empirical study, we show how eye tracking data can be incorporated into choice models and whether the predictive validity can be improved. Finally, both approaches are briefly compared. The paper is concluded with a summarization of main results and suggestions for future research.

Relationship Between Attention and Choice

It can be observed in everyday choice situations that people look longer at things they choose than at things they do not choose (Schotter et al. 2010). When buying new furniture, for example, one would expect that people test and review the features in more detail for those product alternatives they choose afterwards. Pieters and Warlop (1999) stress that most marketing practitioners and academics share the belief that consumers' attention and in-store choice are intimately related. This belief is based on the assumption that the visual attention of a stimulus is a prerequisite that it will be part of the evoked and choice set.

The relationship between attention and choice had not been investigated until Pieters and Warlop (1999) used eye tracking to monitor how consumers make decisions in shelves. They showed that the chosen product receives significantly more attention than the non-chosen ones. They found that three out of the four applied attention measures (fixation duration, number of intra-brand saccades as well as number of inter-brand saccades) increased the likelihood of choice. However, the results from this study are restricted with respect to generalizability because the authors only described their products with different packaging attributes (brand name, pictorial information and ingredient information) in shelf displays. Thus it is an open research question whether the close relation between attention and choice can also be found for more complex products in a typical CBC decision matrix. It can be supposed that the integration of a multitude of attentional processes in complex decision environments produces significant noise in the final choices.

The relationship between attention and choice has been investigated in depth in psychological experiments. Shimojo et al. (2003), for example, showed pictures of faces to

subjects and asked them which of them they find more attractive. By means of eye tracking the authors found that the subjects spend more time on looking at those faces they finally choose in the decision making task. The obvious conclusion from this result was that people look longer at stimuli they like (the respective effect is called "preferential looking"). This result was recently confirmed by Glaholt and Reingold (2009). In their study the attentional focus on stimuli which are finally chosen is significantly longer than the one on non-chosen items.

The main interest of the above-mentioned studies was to investigate the overall relationship between attention and choice. However, in-depth comparisons of attention and choice should also investigate the underlying drivers of choice, such as given by the importances and part-worth-utilities estimated from a choice model. It can be supposed that attributes with higher fixation intensities would also be of higher importance for the decision problem at hand. Moreover, attribute levels with high part-worth utility values should attract a higher amount of attention. A previous study comparing conjoint and process tracing data supports this supposition: Olshavsky and Acito (1980) found a high level of consistency with respect to the importances of attributes between the results of a protocol analysis and a traditional conjoint study. In contrast to this result, Harte, Koele, and van Engelenburg (1996) argue that information display board variables represent other characteristics of the choice process than the importances derived from choice models. The correlation of both measures in CBC is an open research issue.

It has been frequently stated that real choices have little in common with the rational processes that economists have assumed for many years (Adamowicz et al. 2008). Researchers have shown that preferences in many cases are constructed at the time of choice and frequently influenced by contextual factors (Payne, Bettman, and Johnson 1993). This has led to an ongoing debate whether preferences are inherent and stored in the long-term memory or rather context-driven and constructed in the decision situation (Hoeffler and Ariely 1999; Kivetz, Netzer, and Schrift 2008; Simonson 2008). The discussion about preference construction is closely related to the question of attention in choice tasks. Assuming respondents to behave fully rational would imply that individuals attend all relevant information. This is due to the fact that the processing of information would be costless if respondents can process as much information as they need to arrive at the best decision. But since the amount of information that can be processed is limited, attention is a scarce resource. Very often shortcut choice strategies are applied that ignore a lot of information (Todd 2007). In fact, it can be assumed that rational individuals allocate their attention and attend some attributes more than others. They might even ignore certain attribute information. On this account Cameron and DeShazo (2008) recently proposed the introduction of a multiplicative propensity-to-attend parameter in order to arrive at an "attention-corrected choice model" (p. 36).

Analogously, the idea underlying a study by Hensher, Rose, and Greene (2005) was to ask respondents which attributes they did not use when making their choices. Subsequently, the authors included or excluded the respective attributes from the estimation of a mixed logit model based on respondents' information. While their idea of using processing information (i.e. information on attribute ignorance) is quite similar to our approach, the authors did not specifically measure respondents' attention, but only asked them whether they used the attributes consciously. The authors thus concluded that "processing strategies should be built into the estimation of choice data from stated choice studies" (p. 214). Starting from this idea, the present paper also incorporates unconscious attentional processes gathered by eye tracking data into choice models.

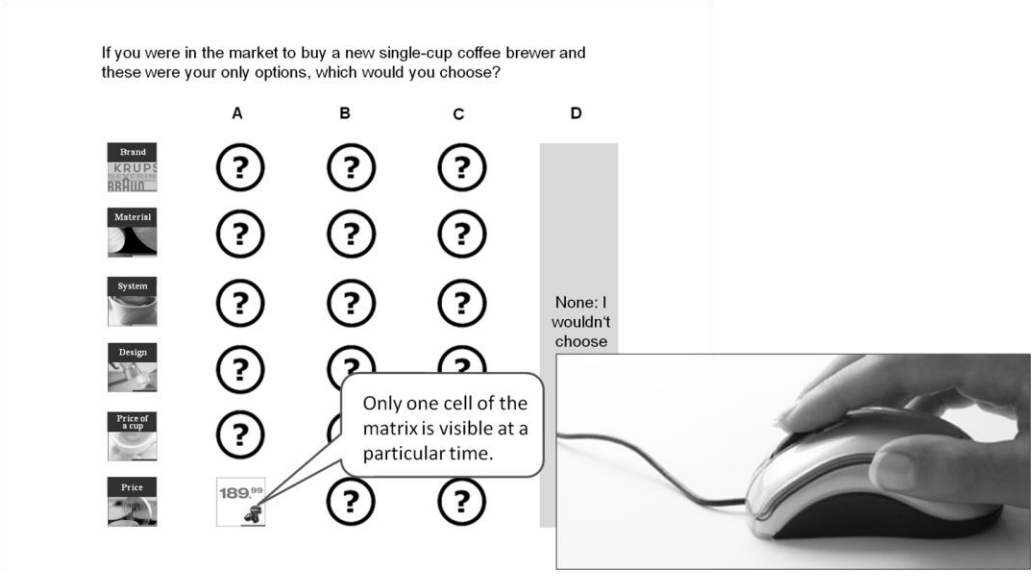
COMPUTERIZED PROCESS TRACING AS A MEANS TO MEASURE ATTENTION

Process tracing techniques can be used to monitor the decision processes of respondents in more detail. Information acquisition data include the amount of information acquired, the sequence of acquisition as well as the time respondents spend on the examination of certain pieces of information. With regard to CBC the main question is how respondents seek information before making a choice and how that information is cognitively integrated.

To date, several process tracing techniques have been developed that are different with respect to the way the data are recorded. In early days, retrospective verbal protocols and information display boards have been used extensively in the behavioral decision making literature (Einhorn and Hogarth 1981; Ford et al. 1989). Nowadays, computerized techniques like Mouselab and eye tracking have replaced manual approaches because information acquisition is more unobtrusive and, thus, more realistic (Lohse and Johnson 1996).

The Mouselab technique is closely related to the research framework of the "adaptive decision maker" (Payne et al. 1993). Based on the idea of information display boards, the choice alternatives are presented in an alternative-by-attribute matrix of covered information cells. The mouse is used as a pointing device to reveal the information. A respondent can access exactly one piece of information at a time by moving the mouse pointer over the matrix cells (see Figure 1). The information is covered again, when the mouse leaves the respective cell. This way, the amount, sequence and duration of information acquisition can be recorded. Hui, Fader, and Bradlow (2007) emphasize that path data retrieved from Mouselab may offer additional insights with respect to the cognitive processes underlying decisions.

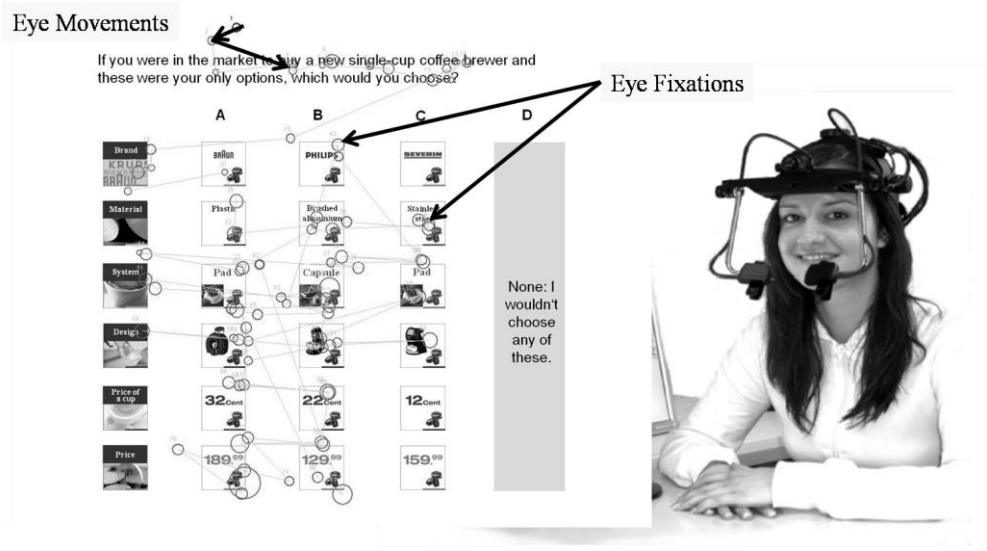
Figure 1
Information acquisition in Mouselab



An even more prominent technique, which is frequently used in psychological experiments, is the recording of respondents' eye movements. Information acquisition is monitored by tracking which matrix cells are fixated with the eyes (see Figure 2). A fixation is the maintaining of the visual gaze on a single location. This means that the spotlight of attention "illuminates" the desired region, for example an attribute level of a decision alternative. During an eye fixation

information is extracted from the perceptual field. The jump of the eye from one fixation to the next is called a saccade. These movements redirect the focus on a new fixation position (Van der Lans, Pieters, and Wedel 2008). During saccades perception is suppressed. According to the "eye-mind hypothesis" (Just and Carpenter 1980) the duration of eye fixations is directly connected to the length of the cognitive process because the major part of visual information is accessed and processed instantaneously (Norman and Schulte-Mecklenbeck 2010). Therefore, eye tracking data do not only measure visual attention, but are an indicator for the amount of cognitive consideration as well.

Figure 2
Information acquisition via eye tracking



Eye tracking has been used in numerous marketing research contexts. Eye movements have been recorded in advertising research (Wedel and Pieters 2000), studies on information search on the Web (Goldberg et al. 2002) and in connection with computer simulated retail shelves (Chandon et al. 2009; Van der Lans et al. 2008). A recent review of eye tracking research in marketing is given by Wedel and Pieters (2007).

It has already been stressed in the literature that a major advantage of eye tracking compared to Mouselab is that the visual attention of respondents is not entirely under cognitive control (Reisen, Hoffrage, and Mast 2008). For example, the respondent may apply unconscious processing like automated scanning routines in a decision situation (Seth et al. 2008). Therefore, the recording of eye movements has been characterized as being more objective because they reflect non-intentional attention (Norman and Schulte-Mecklenbeck 2010). But it has also been mentioned that information display boards like Mouselab might induce a certain kind of information processing (Dieckmann, Dippold, and Dietrich 2009). Accordingly, it has been concluded that eye tracking is better suited for complex decision matrices, i.e. if more attributes are included in a choice task (Reisen et al. 2008).

A clear advantage of Mouselab is that mouse movement recording is easier to implement due to the lower complexity of the measurement itself. Although technical advances have led to eye trackers which are much easier to handle and less expensive than first-generation equipment,

implementation time and effort are still higher for eye tracking studies. However, today's eye tracking technologies can also be used outside laboratories. Heat mounted eye tracking systems (as shown in Figure 2) have a fixed connection between the respondent's head and the cameras. This way a respondent can move freely and investigate the environment, for example at the point-of-purchase (Norman and Schulte-Mecklenbeck 2010). Furthermore, both convenience and accuracy of eye trackings have been improved substantially in recent years due to faster computer processors, among others.

Concluding, from a practitioner's point of view, three difficulties have to be overcome when recording and analyzing eye tracking data: First, it has to be checked whether the eye tracking data can be clearly assigned to the areas of interest. Previous studies have shown that a reliable calibration cannot be achieved for all respondents: For example, recording may be affected if lighting produces shadows or if the test person's eyes are occluded by glasses or makeup. This was also a problem in our study (see below). Second, the experimenter has to determine a certain cut-off level (i.e. duration time in milliseconds) to distinguish fixations from saccades. This makes eye tracking data somehow ambiguous compared to Mouselab data. Third, it has been stressed that the analysis of eye tracking data is challenging because many sources of variation influence the spatiotemporal attention process (Van der Lans et al. 2008).

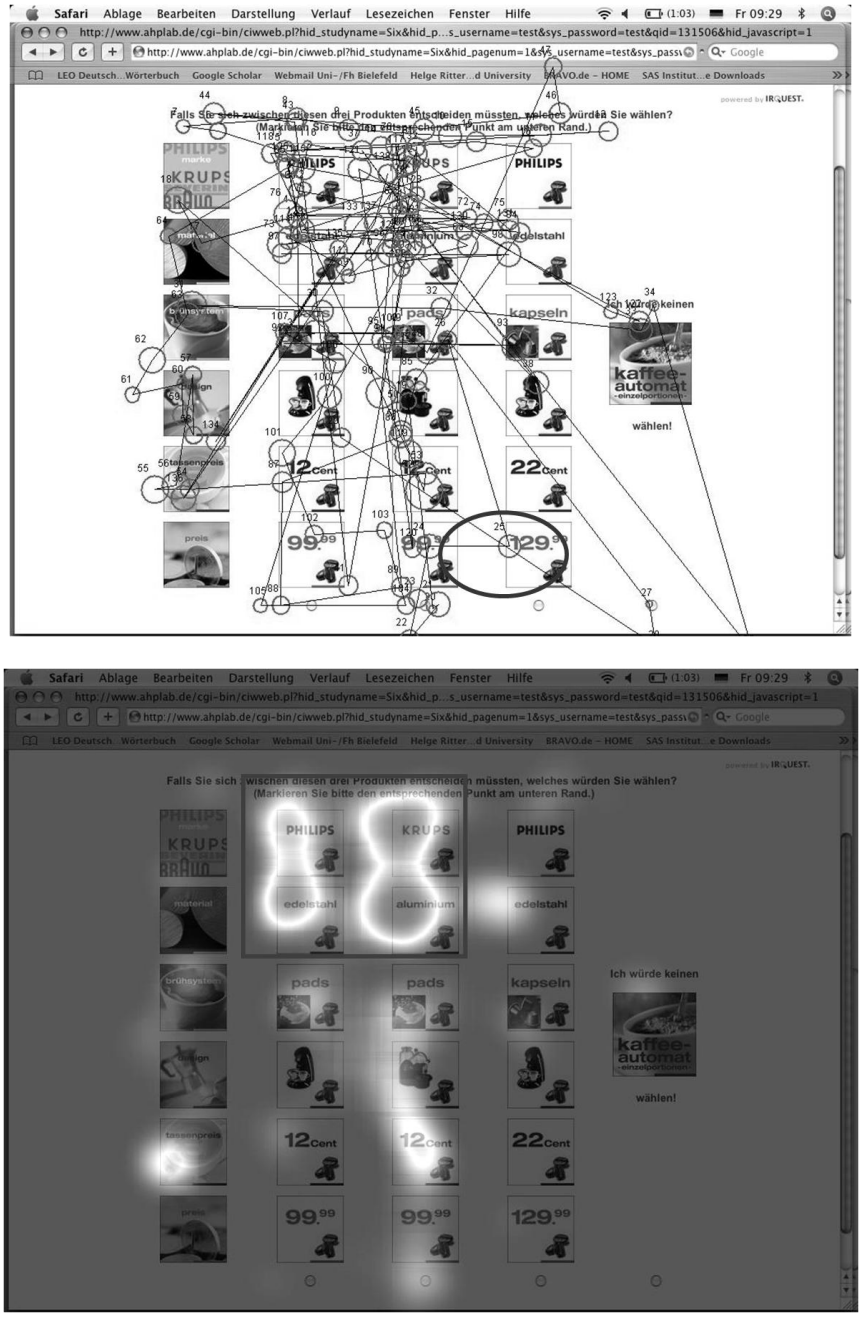
To date, only a few papers have investigated whether process tracing via Mouselab and eye tracking leads to similar results. With respect to choice task data empirical comparisons indicate some differences: Already Lohse and Johnson (1996) could show that respondents need less time and acquire more pieces of information when being eye-tracked. Moreover, respondents had more variable patterns of information search compared to Mouselab. Similar results were reported by Reisen, Hoffrage, and Mast (2008). However, these comparisons did not take place in a marketing research context using CBC.

Visual Inspection of Process Tracing Data

The simplest means to analyze process tracing data is to "replay" the recorded sequences of fixations. In the case of eye tracking this means that the experimenter analyses video recordings of the respondents' eye movements. In doing so, the researcher gets a qualitative idea of how information is processed. Figure 3a depicts all fixations of a respondent in a choice task numbered in consecutive order. As to be seen, the respondent mainly acquires information about the alternatives on the left and in the middle, whereas most of the information concerning the alternative on the right is not evaluated at all. After some initial fixations, the respondent compares the presented alternatives with respect to prices. Due to the fact that the right alternative (129 €) is more expensive than the remaining two (99 €), the respondent seems to exclude the right alternative from further consideration. This can be concluded from the fact that, after the fixation on the price level (129 €), the respondent almost completely stops evaluating the right alternative. The visual inspection of the evaluation process suggests that the higher price of the right alternative proves unacceptable for the respondent.

The fixation data can also be further aggregated and visualized by means of heat maps. In the heat map presented in Figure 3b the frequency of information acquisition is visualized in terms of brightness, i.e. brighter areas are fixated more often. It can be seen that the respondent frequently evaluates the attributes *brand* and *material* in this choice task (cf. the below section on the description of attributes included in the empirical study). In the present decision the respondent obviously trades off these two attributes.

Figure 3
Visualizations of eye tracking data: (a) Fixations pattern suggest unacceptable attribute level (top) / (b) Heat map shows trade-off (bottom)



Quantitative Analysis of Process Tracing Data

In addition to the basic way of recording and visualizing eye movements, fixation data can also be used in quantitative analysis. Researchers developed several process measures which characterize the evaluation process in different ways. Most often these measures are used to determine decision strategies (Wedell and Senter 1997). Transition indices, like the strategy measure (SM, Böckenholt and Hynan 1994) indicate whether a choice task is processed rather

alternative- or rather attribute-wise. This information is most often employed to investigate the shortcuts (decision heuristics) a respondent might have used in order to reduce decision complexity. Another frequently calculated measure is the compensation index (Koele and Westenberg 1995) which quantifies the degree of compensatory search behavior by considering depth of search (i.e. the number of matrix cells opened) and search variability (i.e. the distribution of search efforts across the decision alternatives). Both process measures can help to describe decision processes in more detail. Mintz, Currim, and Jeliazkov (2010), for example, recently investigated how processing patterns affect purchase decisions.

THE SURPLUS OF USING EYE TRACKING IN CBC STUIES – AN EMPIRICAL STUDY

Design of the Empirical Study

The empirical study was conducted using single-cup coffee brewers, mainly for two reasons: First, this kind of coffee brewer has become a high involvement product in Germany in recent years. Second, we assume that most people have rich experience regarding the use and purchase of coffee machines, since coffee is the most favored hot beverage in Germany.

Table 1
Attributes and attribute levels used in the empirical study

Attributes	**Attribute levels**
Brand	Braun, Krups, Philips, Severin
Material	Stainless steel, plastic, brushed aluminum
System	Pad, capsule
Design	Design A, design B, design C, design D
Price of a cup	12 Cents, 22 Cents, 32 Cents
Price (coffee machine)	99 €, 129 €, 159 €, 189 €

A pre-study with 20 respondents was used to identify the six most important attributes by means of the dual questioning approach (Myers and Alpert 1968). Table 1 lists the respective attributes and levels included in the final choice design. The limitation to six attributes is in line with common practice and accounts for methodical requirements of CBC, particularly in view of the fact that information overload could otherwise have impaired the predictive validity. The CBC approach was implemented using Sawtooth Software. We used a standard complete enumeration minimal overlap design to generate twelve choice sets comprising three alternatives for each respondent (Orme 2009).

The computer questionnaire consisted of three parts: In the first part the respondents were surveyed about their consumption of hot beverages. Those respondents who did not drink coffee within the last year were excluded from the survey. Then, all respondents were informed about the attributes and attribute levels describing single-cup coffee brewers by means of textual and pictorial descriptions. Each respondent had to answer three “warm-up choice tasks” previous to the twelve choice tasks being used for CBC estimation. The attribute order in the choice tasks had not been randomized in order to ensure that the presentation of the choice tasks resembled

typical implementations in online consumer research settings. The third part of the survey included some additional questions on the individual socio-demographics.

The Sample

In order to analyze respondents' information processing and acquisition behavior in CBC exercises, computer-aided personal interviews had been conducted under laboratory conditions. In all, 110 (eye tracking) and 91 (Mouselab) adults participated in the studies. Respondent were rewarded for participation.

All respondents had to fill out a CBC computer questionnaire while being eye-tracked with the head-mounted SMI Eye-Link II system. This binocular video-based system measures respondents' gaze position on a 250 Hz video screen (1280x1024 pixels) with an accuracy of 0.5 to 1.0 degree of visual angle. A 9-point calibration routine was executed at the beginning of each measurement. Non-overlapping areas of interest were defined in order to assign the fixations to the different cells of the alternative-by-attribute matrix. The eye-tracking data were checked whether the measured fixations could unambiguously be assigned to the CBC matrix for the 12 choice tasks to be processed by each respondent. This pre-analysis showed that 62 (56.4 percent) interviews could be taken for the subsequent analyses.

Parameter Estimation

Both part-worth utilities and attribute importances were computed at the individual level by applying Hierarchical Bayes (HB) MultiNomial Logit (MNL) estimation (Sawtooth Software Inc. 2009). On average, the HB estimation yields a rather fair goodness of fit. The RLH equals 0.78 for the 62 respondents included in the analysis (eye tracking) and 0.74 (Mouselab) and thus exceeds a naïve model with an RLH of 0.33 by factor 2.3 (eye tracking) and 2.2 (Mouselab), respectively.

Convergence of Eye Movements and Common Findings in Respondent Learning

In their seminal paper "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" Johnson and Orme (1996) identified some typical effects of respondent learning, that is, how respondents' answers change during the course of a CBC interview. The following main effects have been found as the CBC task progresses:

- Respondents speed up in later choice tasks
- Brand becomes less important
- Price becomes more important

By using eye tracking, these effects can be investigated in more detail. More precisely, it can be checked whether the above findings result from shifts in respondents' attention to certain attributes or whether they are particularly based on changes in the individual evaluation of presented information, i.e. changes in the respondents' preference structures.

Figure 4
Total (a, top) and relative (b, bottom) number of fixations in the 12 choice tasks

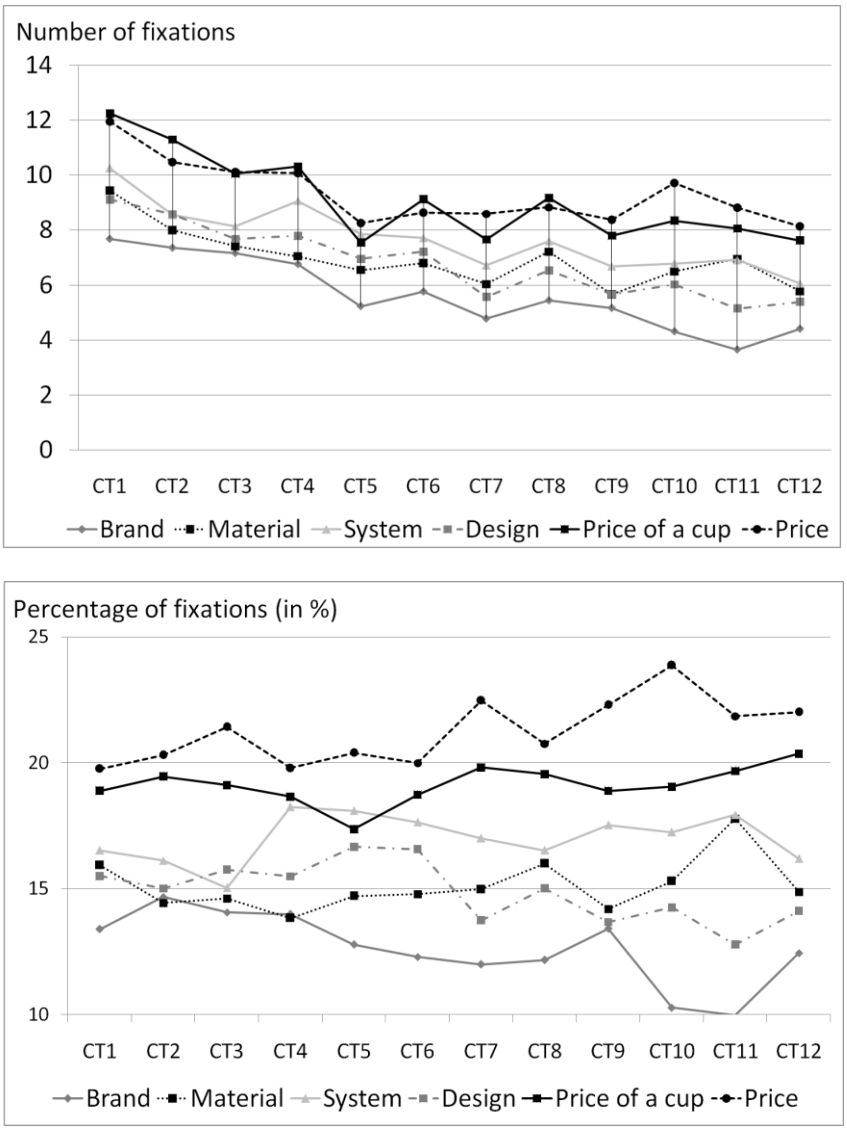


Figure 4 presents the total (a, top) and relative (b, bottom) number of fixations with the twelve choice tasks (CTs). It can be seen that respondents use fewer fixations when the CBC task progresses. On average, the respondents' number of fixations (on the whole screen including the question text and the no-choice option) decreases from 61 in the first choice task to 37 in the last one. However, these differences are rather plausible considering that the respondents get more used to the presented attribute information during the interview. Given the fact that only 18 pieces of information (i.e. 3 alternatives featuring 6 attributes each) are presented on each screen, it seems to be likely that many respondents still pay attention to all displayed attribute information.

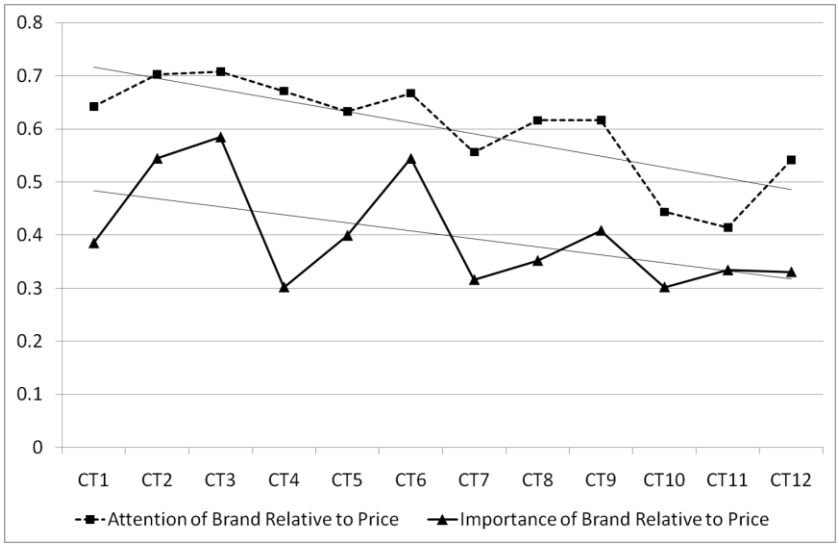
The relative number of fixations, however, draws a different picture. Figure 4b shows a slight shift in attention to the more important attributes included in CBC. For instance, the relative number of fixations on the two monetary attributes machine *price* and *price of a cup* increased

from 19.76 percent (and 18.89 percent, respectively) in the first choice task to 22.01 percent (and 20.36 percent) in the twelfth task. This increase is significant (*price of a cup*, $r = 0.44$, $p < 0.01$; *price*, $r = 0.70$, $p < 0.01$). In contrast to that, the number of fixations on brand decreased significantly (*brand*, $r = -0.74$, $p < 0.01$) during the CBC interview.

In order to further analyze possible effects of shifts in attention on attribute importance we followed the procedure suggested by Johnson and Orme (1996) and carried out aggregate logit estimations for each choice task. Due to the large number of parameters to be estimated the part-worth are relatively unstable when using the data from only one choice task. Nevertheless, the importances should be meaningful. A direct comparison of the aggregate relative importance of attribute *brand* and attribute (coffee machine) *price* against the corresponding relative numbers of fixations shows a clear decline in both parameters (see Figure 5). Obviously, attention is a main driver of choice, although it cannot be assumed to be the only influential factor.

The above results show that studying information processing by means of eye tracking helps to understand usual phenomena emerging during CBC interviews. However, attentional shifts toward certain attributes are not able to fully explain the well documented shifts in preferences. Thus, we have to take a closer look at the relationship between attribute importance and attention in the following.

Figure 5
Shifts in attention and attribute importance for *brand* relative to (coffee machine) *price*

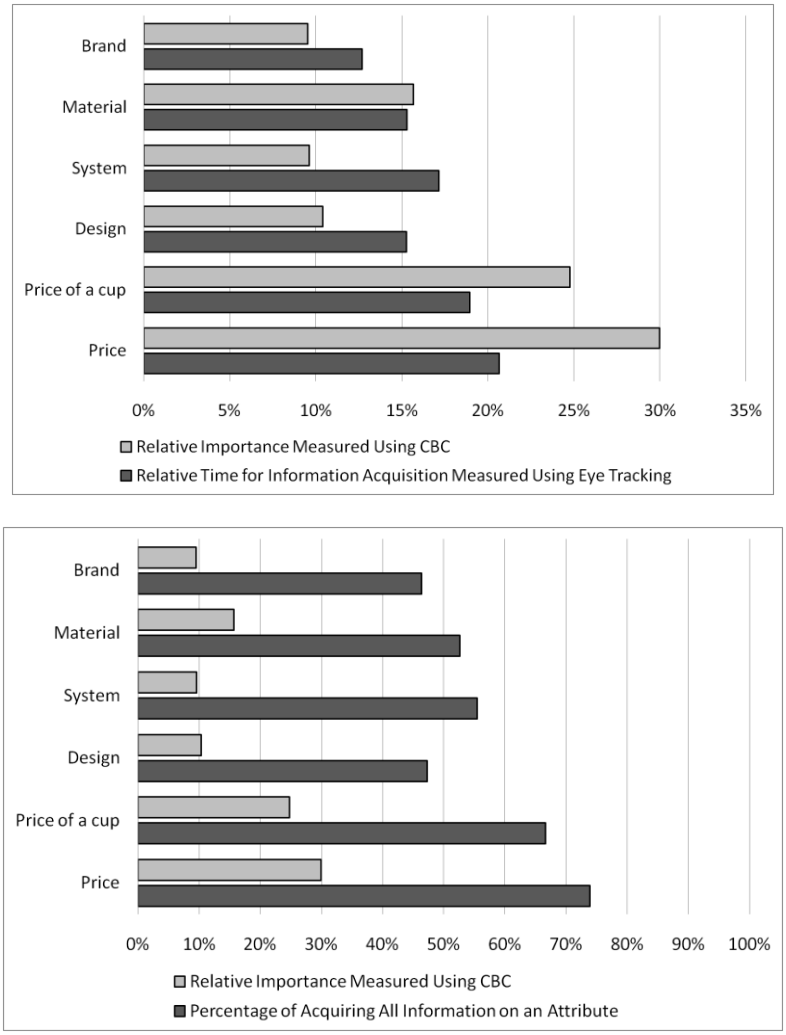


Convergence of Information Processing and CBC Estimates

To better understand the similarities and differences between eye fixations and derived part-worth utilities, we compare the relative number of fixations (subsumed under the term "relevance" in the following) with the aggregate importances of the corresponding attributes (see Figure 6a). As can be taken from the figure, there are substantial differences between attribute importance and the relevance indicated by the relative attention. Noteworthy, easy-to-process information such as the one coming from monetary attributes have lower relevance in comparison to the respective importances, while more complex attributes, such as *design* and *system*, obtain more relevance. Thus, we assume that not only the respondents' preferences but

also the respondents' cognitive efforts to process, encode and store attribute information drives the number of fixations. Nevertheless, both measures are moderately correlated ($r = 0.51$, $p < 0.01$). According to this, attentional focus does at least explain attribute importance to a certain degree although there is substantial noise which cannot be explained by mere information processing effects.

Figure 6
Comparison of attribute importances and relative number of fixations: (a) Importances vs. Relevance (top) / (b) Importances vs. Consideration (bottom)



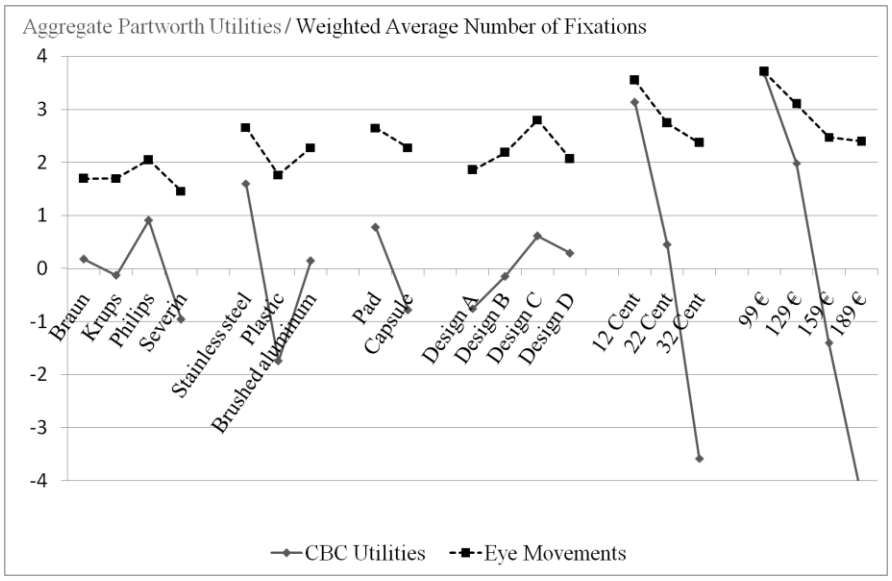
Moreover, some respondents may not process all information on the attribute levels for each alternative in a choice set but rather ignore some of the attribute level information in order to reduce complexity (Scholz, Meißner, and Decker 2010). As outlined above, the ignorance of attribute level information may be caused by non-compensatory decision heuristics, such as elimination by aspects or lexicographic rules. Figure 6b indicates that some attribute information (see *brand* and *design*) is considered for all three choice alternatives in less than 50 percent of the cases. This affirms known findings, that most respondents do not use fully compensatory models in choices (Gilbride and Allenby 2004; Yee et al. 2007).

By means of simple count analysis we are able to compare the aggregate part-worth utilities and the number of fixations devoted to each level (see Figure 7). Because the attributes comprise different numbers of levels, the number of fixations (counts) for each attribute level were weighted with the inverse number of levels of each attribute.

The part-worth utilities and the corresponding attribute level fixation counts follow nearly the same pattern. That is, both measures derive the same within attribute ranking. Using Pearson's correlation coefficient we found a substantial positive correlation between the part-worth utilities and the number of fixations ($r = 0.58$, $p < 0.01$).

Concluding, attribute attention and derived consumer preferences (i.e. part-worth utilities and importances) may investigate consumer preference formation from different angles. While some aspects of the respondents' attentional focus may be included in the estimated part-worth utilities, other aspects, such as cognitive efforts and decision heuristics, may not be fully captured in the estimated compensatory choice model. For this reason, we tried to amend the traditional HB-MNL approach by integrating respondents' individual information processing behavior.

Figure 7
Comparison of aggregate part-worth utilities and weighted average number of fixations



INTEGRATING INFORMATION PROCESSING AND CHOICE – A JOINT MNL APPROACH

While above eye-tracking data and HB parameter estimates are directly contrasted, eye-tracking information may also improve the estimation of utilities. We therefore consider the following two ways of amending simple choices by means of individual information processing behavior:

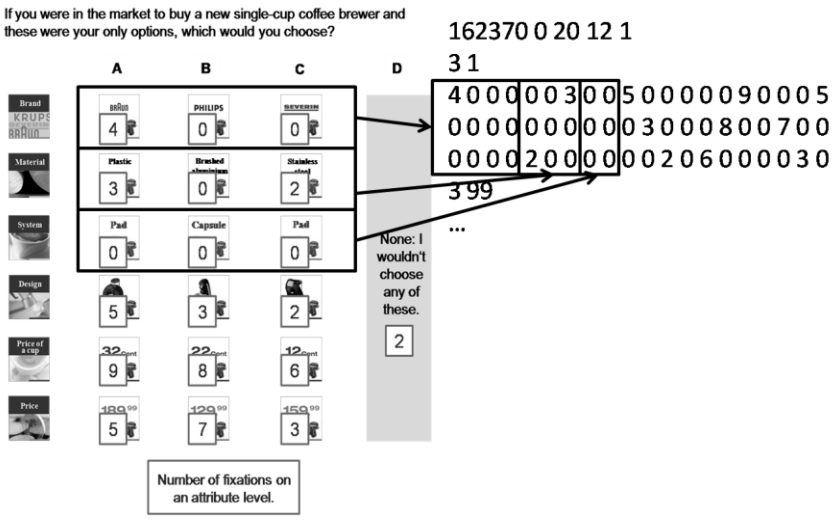
Consideration-based HB-MNL: Inclusion of the non-consideration of attribute levels in the HB-MNL model

Attention-based HB-MNL: Inclusion of the number of fixations for each attribute level in the HB-MNL model

The consideration-based HB-MNL model is pretty straightforward. Based on the number of fixations, attribute levels that are not considered in the choice task are omitted from the parameter estimation. That is, completely ignored attribute level information is set to a missing value in the .cho-file of Sawtooth Software (Consideration-based HB-MNL).

Not only consideration vs. ignorance information may be used to amend the choice model but rather the number of fixations on each attribute may also provide valuable information for better choice predictions. Thus, we include this additional information in the design matrix used for estimating the parameters. In order to do so we manually generated a .cho-file as outlined in Figure 8.

Figure 8
Transferring Fixation Data into .cho-file Structure



Here, each level is coded as a linear attribute comprising the number of fixations. In doing so, the number of parameters increases slightly from (4-1)+(3-1)+(2-1)+(4-1)+(3-1)+(4-1) = 14 to 20. Noteworthy, the resulting parameter estimates cannot be interpreted in the same manner as usual part-worth utilities. Rather, the estimated parameters account for the increase or decrease in overall utility of an alternative when a certain attribute level is fixated. Therefore, we refer to these parameters as information utilities. Of course, the number of fixations is only measurable under laboratory conditions. The attentional focus on attribute level information in real-world purchase decisions is unknown. Accordingly, we have to anticipate the number of fixations in the holdout tasks by means of the eye-tracking data collected in the CBC interview. In the following, we will consider three different assumptions: (a) Each respondent pays the same attention to all attribute levels in the holdout tasks. We refer to this model as attention-based MNL1. (b) The respondent's attentional focus in the holdout tasks is equal to his/her average information processing behavior in the CBC task. In this case, it is straightforward to determine the mean number of fixations for each attribute level, when presented in the choice task at hand. This model is referred to as attention-based MNL2. (c) We know the number of fixations for each attribute level in the holdout tasks. Please recall that the three variances do not differ in

parameter estimations, but in the assumptions on holdout information processing. This model is called attention-based MNL3.

The above estimation models have been compared to the traditional HB-MNL approach. We used the first 10 choice tasks for the estimation of the parameters and the last two choice tasks as holdouts. Table 2 outlines both the goodness-of-fit (RLH) as well as the first choice hit rates for the two holdout choice tasks.

A comparison of the different estimation models shows that the inclusion of attribute ignorance does not affect model fit but substantially improves holdout prediction. At the same time, the attention-based MNL models yield an impressive increase in model fit. Accordingly, the inclusion of information-processing amends the traditional HB-MNL approach. The first-choice hit-rates, however, also show that the inclusion of the attentional focus is not helpful when no specific assumptions are available on how respondents process product alternatives in real-world situations. The attention-based MNL1, which assumes that all attribute level information is processed uniformly, yields exactly the same hit-rate as the traditional HB-MNL model. However, when justifiable assumptions on respondents' attentional focus can be derived, the holdout predictions substantially increase, as can be seen from the hit rates of attention-based MNL2 and MNL3. Due to the small sample size this improvements do not prove statistically significant (MNL2: $p = 0.43$; MNL3: $p = 0.19$).

Table 2
Goodness-of-fit and first choice hit-rates on holdout tasks

Model	RLH	First-choice hit-rate (in percent)
HB-MNL	0.78	58.87
Consideration-based MNL	0.78	61.29
Attention-based MNL1	0.93	58.87
Attention-based MNL2	0.93	63.71
Attention-based MNL3	0.93	66.94

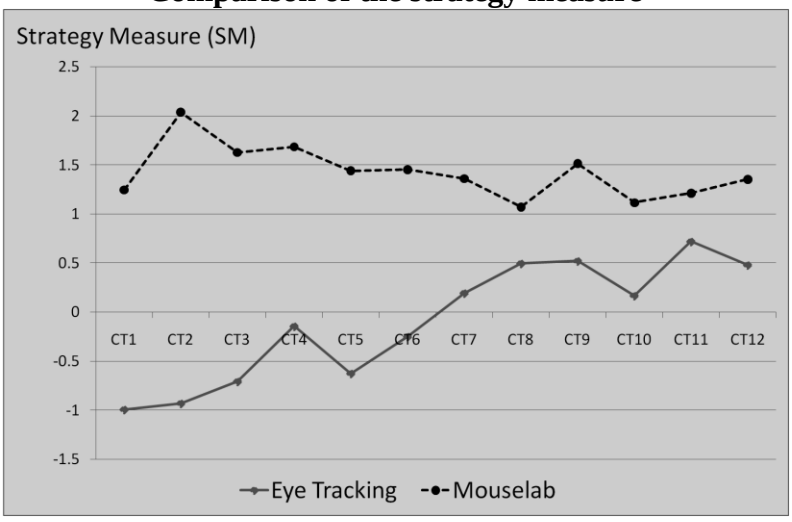
USEFULNESS OF THE MOUSELAB TECHNIQUE

To statistically check whether process tracing information from Mouselab was substantially different from eye tracking, we compared both approaches by means of key measures: With respect to the *number of fixations* both techniques came to significantly different results ($t = 1.652$; $FG = 147$; $p = 0.101$). In case of eye tracking respondents fixated $\bar{x} = 32.226$ matrix cells on average (with standard deviation $\sigma = 10.703$), whereas only $\bar{x} = 28.993$ ($\sigma = 13.425$) cells were clicked in the Mouselab setting. Significant differences were also found for the average *duration of the choice task*. With $\bar{x} = 19.901$ ($\sigma = 7.588$) seconds in Mouselab the mean duration was about 47 percent longer compared to eye tracking with $\bar{x} = 13.546$ ($\sigma = 7.588$). Moreover, with 78.506 percent of matrix cells having been fixated at least once the mean *depth of search* was significantly ($U = 2774$; $Z = -0.175$; $p = 0.861$) higher for eye tracking than for

Mouselab (75.639 percent). Most of the results concerning the referred three measures were in line with finding from previous studies (Lohse and Johnson 1996; Reisen et al. 2008).

HB parameter estimates were also calculated for the choice data conducted with Mouselab. The mean RLH was 0.74. Compared to RLH = 0.78 for the eye tracking sample the internal validity is slightly worse, but this also shows that Mouselab does not necessarily affect the internal validity. This result is supported by the fact that *U*-tests on the differences of the part-worth utilities unveiled that only 3 out of 20 parameters showed significant differences ($\alpha = 0.05$). Moreover, in both conditions the percentage of respondents using the no-choice option was not significantly ($\chi^2_{emp} = 0.830$; FG = 1; $p = 0.362$) different (eye tracking: 17.88 percent; Mouselab 17.58 percent). To put it in a nutshell, these results show that Mouselab should only marginally influence corresponding preference measurements.

Figure 9
Comparison of the strategy measure



Finally, both process tracing approaches were compared with respect to the strategy measure (SM) which indicates how information is processed in choice tasks (CTs). As Figure 9 shows, respondents (on average) switch from a more attribute-wise (indicated by negative values) to a more alternative-wise (indicated by positive values) processing during the course of the CBC interviews when being eye-tracked. This suggests the conclusion that respondents tend to evaluate decision alternatives in choice tasks more holistically, maybe because they are more familiar with the decision situation. In contrast to that information processing tends to be alternative-wise during the whole CBC interviews under Mouselab. This supports the assumption that information processing is biased by Mouselab towards an alternative-wise evaluation of choice tasks. An obvious explanation for this might be the fact that Mouselab reduces the speed of comparison and hinders intuitive processing (see above). We therefore conclude that it is better to use eye tracking when focusing on the description of acquisition behavior.

CONCLUSIONS

This paper investigated how respondents process information in CBC. Eye tracking and Mouselab were used to record information acquisition and investigate the attentional processes in

a typical CBC setting. It has been shown that process tracing data can be used to qualitatively analyze how respondents approach purchase decisions, cross-check and validate CBC utilities and importances, check whether the relevance of attributes and the type of information processing changes during interviews, and improve the predictive validity of choice models.

The first aim of our research was to investigate the relationship between attention and choice. We started with a literature review of previous findings. Most of the studies published to date indicate a strong connection of these two process-steps. In order to further analyze this relationship, we compared CBC and eye tracking data. The results showed that the attentional focus partially explains attribute importance. Moreover, part-worth utilities proved to be positively correlated with the amount of information being processed on the attribute levels. Next, we analyzed the possible effects of shifts in attention on attribute importance. We could replicate the effect shown by Johnson and Orme (1996) that the relative importance of the brand compared to the price decreases during CBC interviews. This relative decrease in importance was mirrored by a relative decrease of attention of brand relative to price. Obviously, attentional shifts toward certain attributes at least partially explain the well-documented shifts in preferences.

The second aim of the study was to test whether data from the attentional process can be used to improve choice models. We therefore included eye tracking data in the standard Hierarchical Bayes MultiNomial Logit model. The comparison of different estimation models showed that the inclusion of attribute ignorance can improve first choice hit rates, but does not necessarily improve model fit. However, when including the average amount of attention on attribute levels in the model, this strongly increased model fit. Moreover, hit rates increased if the attentional focus in choice tasks was used as an additional model input. In the present case, we used the usual concept of holdout choice tasks which had been designed in the same way as the choice tasks used for parameter estimation. Accordingly, the assumption of similar information processing seems to be justifiable. Whether the same applies if respondents are faced with real-world stimuli, which do not follow the usual (verbal) full-profile stimulus design usually applied in CBC (see Figure 2), is at least questionable. Nevertheless, this – once more – stresses an often discussed issue concerning choice experiments, namely the question how far the stimulus representation in CBC leads to differences in information processing. The above results suggest the supposition that both information processing and the final choices are at least partially driven by attention.

The third aim of our study was to compare eye tracking with the easier to handle Mouselab technique. The main result of this comparison is that Mouselab does not impair the outcome of preference measurement, but information processing is biased toward an alternative-wise evaluation of the choice tasks. This means that eye tracking should not be replaced by Mouselab if the focus of a study is on unobtrusively determining information processing strategies. However, due to the fact that Mouselab does not affect the quality of preference measurements and can be easily implemented in web surveys, it can still be valuable alternative.

Especially for complex products the identification of those attributes which are really relevant in purchase decision making is a major challenge. Since this research has shown that process tracing approaches are useful tools to identify whether a certain attribute is considered in a choice task or not, future research should investigate how data from the attentional step can be used for the selection of attributes in preference measurement.

Of course, this study is not free of limitations. This particularly concerns the generalizability of the empirical outcomes. Due to the fact that eye-tracking respondents in the laboratory is relatively time consuming, the final sample consisted of only 62 respondents the data of whom could be used in the MNL-estimations. Although the hit rates achieved a satisfactory level, further replications of these analyses using larger sample sizes and different products are necessary.

REFERENCES

Adamowicz, Wiktor, David Bunch, Trudy Cameron, Benedict Dellaert, Michael Hanneman, Michael Keane, Jordan Louviere, Robert Meyer, Thomas Steenburgh, and Joffre Swait (2008), "Behavioral Frontiers in Choice Modeling," *Marketing Letters*, 19 (3), 215-228.

Böckenholt, Ulf and Linda S. Hynan (1994), "Caveats on a Process-Tracing Measure and a Remedy," *Journal of Behavioral Decision Making*, 7 (2), 103-117.

Bradlow, Eric T. (2005), "Current Issues and a Wish List for Conjoint Analysis," *Applied Stochastic Models in Business and Industry*, 21 (4-5), 319-323.

Cameron, Trudy Ann and J. R. DeShazo (2008), "Differential Attention to Attributes in Utility-Theoretic Choice Models," in *Working paper*.

Chandon, Pierre, J. Wesley Hutchinson, Eric T. Bradlow, and Scott H. Yount (2009), "Does in-Store Marketing Work? Effects of the Number and Position of Shelf Facings on Brand Attention and Evaluation at the Point of Purchase," *Journal of Marketing*, 73 (6), 1-17.

Dieckmann, Anja, Katrin Dippold, and Holger Dietrich (2009), "Compensatory Versus Noncompensatory Models for Predicting Consumer Preferences," *Judgment and Decision Making*, 4 (3), 200-213.

Einhorn, Hillel J. and Robin M. Hogarth (1981), "Behavioral Decision Theory: Processes of Judgment and Choice," *Annual Review of Psychology*, 32, 53-88.

Ford, J. Kevin, Neal Schmitt, Susan L. Schechtman, Brian M. Hults, and Mary L. Doherty (1989), "Process Tracing Methods: Contributions, Problems, and Neglected Research Questions," *Organizational Behavior and Human Decision Processes*, 43 (1), 75-117.

Gilbride, Timothy J. and Greg M. Allenby (2004), "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23 (3), 391-406.

Glaholt, Mackenzie G. and Eyal M. Reingold (2009), "Stimulus Exposure and Gaze Bias: A Further Test of the Gaze Cascade Model," *Attention, Perception, & Psychophysics*, 71 (3), 445-450.

Goldberg, Joseph H., Mark J. Stimson, Marion Lewenstein, Neil Scott, and Anna M. Wichansky (2002), "Eye Tracking in Web Search Tasks: Design Implications," in *Proceedings of the Symposiumon Eye Tracking Research & Applications*: ACM Press, 51-58.

Harte, Johanna M., Pieter Koele, and Gijsbert van Engelenburg (1996), "Estimation of Attribute Weights in a Multiattribute Choice Situation," *Acta Psychologica*, 93 (1-3), 37-55.

Hensher, David A., John Rose, and William H. Greene (2005), "The Implications on Willingness to Pay of Respondents Ignoring Specific Attributes," *Transportation*, 32 (3), 203-222.

Hoeffler, Steve and Dan Ariely (1999), "Constructing Stable Preferences," *Journal of Consumer Psychology*, 8 (2), 113-139.

Hui, Sam K., Peter S. Fader, and Eric T. Bradlow (2009), "Path Data in Marketing: An Integrative Framework and Prospectus for Model Building," *Marketing Science*, 28 (2), 320-335.

Jasper, J. D. and Jennifer Shapiro (2002), "Mousetrace: A Better Mousetrap for Catching Decision Processes," *Behavior Research Methods, Instruments, & Computers*, 34 (3), 364-374.

Johnson, Richard M. and Bryan K. Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?," in *Technical Paper Series*, Sequim: Sawtooth Software Inc., http://www.sawtoothsoftware.com/download/techpap/howmanyq.pdf.

Just, Marcel Adam and Patricia A. Carpenter (1980), "A Theory of Reading: From Eye Fixations to Comprehension," *Psychological Review*, 87 (4), 329-354.

Kivetz, Ran, Oded Netzer, and Rom Y. Schrift (2008), "The Synthesis of Preference: Bridging Behavioral Decision Research and Marketing Science," *Journal of Consumer Psychology*, 18 (3), 179-186.

Koele, Pieter and Mirjam R. M. Westenberg (1995), "A Compensation Index for Multiattribute Decision Strategies," *The Psychonomic Bulletin and Review*, 2 (3), 398-402.

Logan, Gordon D. and N. Jane Zbrodoff (1999), "Selection for Cognition: Cognitive Constraints on Visual Spatial Attention," *Visual Cognition*, 6 (1), 55-81.

Lohse, Gerald L. and Eric J. Johnson (1996), "A Comparison of Two Process Tracing Methods on Choice Tasks," *Organizational Behavior and Human Decision Processes*, 68 (1), 28-43.

Mintz, Ofer, Imram S. Currim, and Ivan Jeliazkov (2010), "Consumer Search and Propensity to Buy," http://www.economics.uci.edu/~ivan/MCJ2010.pdf.

Myers, James H. and Mark I. Alpert (1968), "Determinant Buying Attitudes: Meaning and Measurement," *Journal of Marketing*, 32 (4), 13-20.

Netzer, Oded, Olivier Toubia, Eric Bradlow, Ely Dahan, Theodoros Evgeniou, Fred Feinberg, Eleanor Feit, Sam Hui, Joseph Johnson, John Liechty, James Orlin, and Vithala R. Rao (2008), "Beyond Conjoint Analysis: Advances in Preference Measurement," *Marketing Letters*, 19 (3), 337-354.

Norman, Elisabeth and Michael Schulte-Mecklenbeck (2010), "Take a Quick Click at That! Mouselab and Eye-Tracking as Tools to Measure Intuition," in *Tracing Intuition: Recent Methods in Measuring Intuitive and Deliberate Processes in Decision Making*, ed. A. Glöckner and C. L. M. Wittemann, London: Psychology Press / Routledge, 24-44.

Olshavsky, Richard and Franklin Acito (1980), "An Information Processing Probe into Conjoint Analysis," *Decision Sciences*, 11 (3), 451-470.

Orme, Bryan (2009), "Fine-Tuning CBC and Adaptive CBC Questionnaires," in *Technical Paper Series*, Sequim: Sawtooth Software Inc., http://www.sawtoothsoftware.com/download/techpap/finetune.pdf.

Payne, John W., James R. Bettman, and Eric J. Johnson (1993), *The Adaptive Decision Maker*, Cambridge: Cambridge University Press.

Pieters, Rik and Luk Warlop (1999), "Visual Attention During Brand Choice: The Impact of Time Pressure and Task Motivation," *International Journal of Research in Marketing*, 16 (1), 1-16.

Reisen, Nils, Ulrich Hoffrage, and Fred W. Mast (2008), "Identifying Decision Strategies in a Consumer Choice Situation," *Judgment and Decision Making*, 3 (8), 641-658.

Sawtooth Software Inc. (2009), "The CBC/HB System for Hierarchical Bayes Estimation Version 5.0 Technical Paper," in *Technical Paper Series*, Sequim: Sawtooth Software Inc., http://www.sawtoothsoftware.com/download/techpap/hbtech.pdf.

Scholz, Sören W., Martin Meißner, and Reinhold Decker (2010), "Measuring Consumer Preferences for Complex Products: A Compositional Approach Based on Paired Comparisons," *Journal of Marketing Research*, 47 (4), 685-698.

Schotter, Elizabeth R., Raymond W. Berry, Craig R. M. McKenzie, and Keith Rayner (2010), "Gaze Bias: Selective Encoding and Liking Effects," *Visual Cognition*, 18 (8), 1113-1132.

Seth, Anil K., Zoltán Dienes, Axel Cleeremans, Morten Overgaard, and Luiz Pessoa (2008), "Measuring Consciousness: Relating Behavioural and Neurophysiological Approaches," *Trends in Cognitive Sciences*, 12 (8), 314-321.

Shimojo, Shinsuke, Claudiu Simion, Eiko Shimojo, and Christian Scheier (2003), "Gaze Bias Both Reflects and Influences Preference," *Nature Neuroscience*, 6 (12), 1317-1322.

Simonson, Itamar (2008), "Will I Like a “Medium” Pillow? Another Look at Constructed and Inherent Preferences," *Journal of Consumer Psychology*, 18 (3), 155-169.

Todd, Peter M. (2007), "How Much Information Do We Need?," *European Journal of Operational Research*, 177 (3), 1317-1332.

Van der Lans, Ralf, Rik Pieters, and Michel Wedel (2008), "Eye-Movement Analysis of Search Effectiveness," *Journal of the American Statistical Association*, 103 (482), 452-461.

Wedel, Michel and Rik Pieters (2000), "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings," *Marketing Science*, 19 (4), 297-312.

--- (2007), "A Review of Eye-Tracking Research in Marketing," in *Review of Marketing Research*, Vol. 4, ed. Naresh K. Malhotra, New York: M. E. Sharp Inc., 123-147.

Wedell, Douglas H. and Stuart M. Senter (1997), "Looking and Weighting in Judgment and Choice," *Organizational Behavior and Human Decision Processes*, 70 (1), 41-64.

Yee, Michael, Ely Dahan, John R. Hauser, and James Orlin (2007), "Greedoid-Based Noncompensatory Inference," *Marketing Science*, 26 (4), 532-549.

THE VALUE OF CONJOINT ANALYSIS IN HEALTH CARE FOR THE INDIVIDUAL PATIENT

LIANA FRAENKEL
YALE UNIVERSITY SCHOOL OF MEDICINE,
VA CONNECTICUT HEALTHCARE SYSTEM

VARIABILITY IN THE DELIVERY OF HEALTHCARE SERVICES

There are two types of variability in medicine: unwarranted and warranted. Unwarranted variability refers to variable delivery of healthcare services that are not explained by medical need. For example, elective back surgery and caesarian section rates are much higher in some states than others. This pattern is largely driven by differences in physician preferences. Practice guidelines and performances measures are meant to decrease unwarranted variability and promote consistent high quality care.

In contrast, warranted variability is due to differences in patient preferences. Warranted variability is most apparent in value-sensitive decisions such as colorectal cancer screening, treatment for prostate cancer, as well as treatment for arthritis and other chronic diseases. In these cases, variability is expected, and treatment decisions should be based on both physician judgment and explicitly derived patient preferences. This process of shared decision-making not only adheres to the principles of informed consent and patient autonomy, but also increases patient satisfaction, and may improve compliance with treatment and health outcomes (1-3).

CONCEPTUAL MODEL

Over the past two decades, the physician-patient relationship has shifted from a paternalistic model towards a shared decision-making model, in which patients are encouraged to play an active role in decisions concerning their health care. This model rests on the assumption that patients want to be fully informed about issues related to their health care, but recognizes that patients' desire to actually participate in decision-making is variable. This assumption is supported by many studies examining patients' preferences for information and participation in medical decision-making. We and others have found that the vast majority of patients, regardless of disease type and demographic characteristics, want to be fully informed of all available treatment options and their related risks (4). Furthermore, research has shown that provision of information improves compliance and health outcomes and does not increase patient anxiety (1-3).

In contrast, patient preferences for participating in decision-making appear to be more variable, with younger patients tending to prefer a more active role and older patients a more passive role (5). However, elicitation of individual patient preferences is an essential component of decision-making for all patients, regardless of the amount of control they want to have over the final decision. This view is in keeping with a growing literature demonstrating that physicians' and patients' values often diverge and that physicians are poor predictors of individual patient values (6, 7). Therefore, decisions involving personal values should be based on individual patient values as well as physician judgment. This is also true for patients who

prefer a more passive role, since physicians can only decide on the "best" treatment plan for each patient if they have a clear understanding of how that particular individual values specific trade-offs.

Measuring Patient Preferences

One of the main reasons underlying the lack of shared decision-making in clinical practice is the paucity of tools available to help providers effectively communicate complex medical information to their patients in a manageable way. Patient preferences are most commonly measured using one or a combination of the following three techniques: standard gamble, time trade-off, or rating scales. All three modalities quantify preferences for defined health states that can then be incorporated into decision models to calculate quality-adjusted outcomes (such as quality adjusted life years). There are, however, numerous limitations associated with these techniques, including poor inter-method agreement, susceptibility to biases, unproven predictive validity, and the degree of difficulty associated with the standard gamble and time trade-off tasks (8-11). Moreover, these methods are not accurate enough to facilitate decision-making at the individual patient level, nor do they provide insights into the reasons underlying individual patient preferences for a particular treatment option or intervention. In addition, a subgroup of patients appears unwilling to trade off years of life for religious and other reasons. Other investigators have used probability trade-offs tasks to quantify patient preferences. While generally easier to perform than the standard gamble and time trade-off tasks, probability trade-off tasks are based solely on patients' ability to comprehend small differences in probability estimates and are limited in the number of treatment characteristics they can evaluate.

Conjoint Analysis

Conjoint analysis (CA) has many properties which make it a valuable tool to elicit patient preferences and facilitate medical decision-making. It can be designed to ensure that patients are made aware of all essential information related to appropriate treatment options, and therefore should improve patient knowledge and informed consent. In addition, CA questionnaires may be easily formatted to present individualized estimates of risks and benefits. CA has the potential to improve the quality of decisions by making the trade-offs between competing options explicit. This is of direct clinical relevance since choices based on explicit trade-offs are less likely to be influenced by heuristics (errors in reasoning) which can lead to poor decisions. In addition, CA can be used to examine the amount of importance respondents place on specific treatment characteristics which enables physicians to gain insight into the reasons underlying their patients' preferences, tailor discussions to address individual patient's concerns, and ensure that decisions are made based on accurate expectations.

ACA may be of particular value because it is interactive, and therefore more efficient than other techniques thus allowing a large number of attributes to be evaluated without resulting in information overload or respondent fatigue. This is an important advantage, since complex treatment decisions often require multiple trade-offs between competing risks and benefits. ACA also enables patients to receive immediate feedback, an important feature given that our goal is to develop a tool that can be used in clinical settings. Like all CA approaches, ACA provides simulation capability. This feature allows the investigator to assess the impact of varying specific treatment characteristics on choice. For example, researchers can determine how much benefit

patients require before accepting the risk of drug toxicity, whether decreasing the burden or inconveniences of therapy might increase patient acceptance of treatment, or how varying co-pays influence treatment choices.

Quality of Decision-Making

ACA also promotes several of the steps recommended for optimal decision-making. There is a vast literature, based largely on economic principles, describing how people should make decisions in order to maximize benefits. Empirical research has, however, established that people do not consistently make decisions to maximize expected benefits (12-14). Instead, people often rely on heuristics, or simplifying tactics, to facilitate decision-making. This is especially true in cases involving decision-making under uncertainty, which typifies most difficult heath care decisions. While heuristics may simplify the decision-making process, this approach often results in inadequate consideration of available alternatives and poor decisions. To improve the quality of decision-making, social scientists advocate the following steps:

1. Search for available options.
2. Carefully weigh expected benefits and risks associated with each alternative action.
3. Search for new information to further evaluate alternatives.
4. Take into account new information, even if it contradicts initial views.
5. Reexamine positive and negative consequences of each alternative.
6. Make plans to implement the decision.
7. Make contingency plans to deal with risks in case they materialize.

Research shows that failure to adhere to these principles results in poor decision-making. ACA encourages patients to consider all relevant trade-offs involved in complex decisions, and therefore may promote the actions described in steps #2 through #5.

In the next section of this paper we describe some practical examples of how we have used ACA and other CA approaches in healthcare applications.

Lupus

Lupus is a serious disease affecting young women of childbearing age. Our first study exploring the value of CA in medical decision making, compared women's treatment preferences with the standard of care for patients with kidney disease due to lupus (lupus nephritis) (15). Standard care for lupus nephritis at the time the study was done included treatment with cyclophosphamide, a fairly potent drug with a significant risk of side effects including a risk of infertility. Azathioprine, a less effective but safer option, was considered a second-line option for patients with contraindications to cyclophosphamide. We hypothesized that women's preferences for these two drugs would vary and that a substantial number of women of childbearing age wanting to have children would prefer azathioprine over cyclophosphamide. To test this hypothesis we administered an ACA survey to 103 women with lupus. Attributes and levels were chosen to represent the range of possible risks and benefits related to both treatment options published in clinical trials. We found that of the nine medication characteristics studied, efficacy

and risk for infection had the greatest impact on preference. As predicted, women wanting more children were less likely to choose cyclophosphamide compared with their counterparts (56% vs.80%). Although we originally hypothesized that patient preferences would differ from standard care, we were surprised to find that only 56% of women wanting more children preferred cyclophosphamide, even when it conferred the maximum benefit reported in the literature and a low probability of toxicity. Reducing the risk for premature ovarian failure by 50% increased the percentage of women preferring cyclophosphamide by only 8%, suggesting that a significant number of premenopausal women wanting more children are unwilling to accept even the smallest risk of infertility. This study demonstrated that ACA is a feasible and valuable method of evaluating patient treatment preferences.

Osteoarthritis

In a separate study we administered an ACA survey to 100 patients with knee arthritis (16). Explicit elicitation of patient preferences is of particular importance in the treatment of patients with knee arthritis, because pharmacologic options have relatively modest efficacy and differ significantly with respect to their risk of drug toxicity and cost. In this study, we found that many older patients with knee arthritis are willing to forego treatment effectiveness for a lower risk of adverse effects. Anti-inflammatory drugs, the most widely prescribed medication for patients with arthritis, was the least preferred therapeutic option across almost all simulations. The magnitude of the discrepancy between patient preferences in this study and the widespread use of nonselective anti-inflammatory drugs raises important questions about how patient preferences are elicited and how treatment decisions for osteoarthritis are made in clinical practice.

We subsequently conducted a pilot randomized controlled clinical trial to examine the feasibility of using ACA to elicit preferences and improve decision making in clinical practice (17). In this study patients with knee pain were randomized to receive an information pamphlet or to perform an ACA task. The latter was designed to elicit preferences based on patient tradeoffs for route of administration, benefits, and side effects of commonly used treatment options for knee pain. After performing the task, participants were given a printed handout illustrating their preferences. 87 patients were randomized. We found that decisional self-efficacy, preparedness to participate in decision-making, and arthritis self-efficacy were greater in participants randomized to the intervention arm compared to those receiving the information pamphlet ($p < 0.05$ for all comparisons) (17). These results indicate that participants using a tool designed to increase patient awareness of choice and evaluate the tradeoffs related to available treatment options were more confident in their ability to obtain information about available treatment options, were better prepared to participate in their visit, and had better arthritis related self-efficacy compared to patients receiving an information pamphlet. The results of this pilot study justify future large-scale trials to determine the effectiveness of similar interventions.

Rheumatoid Arthritis

ACA may also be a valuable method of gaining insights underlying variability in patient preferences. For example, racial disparities have been noted in outcomes and the delivery of healthcare services in chronic disease. In many studies, Black patients tend to be less likely than their White counterparts to choose invasive procedures or aggressive treatment. Whether variability in treatment preferences accounts for this difference is not known. To examine this

possibility we elicited treatment preferences using ACA for aggressive therapy in 136 patients with rheumatoid arthritis who identified themselves as being Black or White (18). In unadjusted analysis, 51% of White participants preferred aggressive therapy compared to 16% of Blacks ($p<0.0001$). Race remained the strongest predictor of aggressive therapy after adjusting for relevant co-variates.

We also created a variable representing the ratio of the importance that patients attach to overall benefit (average of values for all benefits) versus overall risk (average of values for all risks) (19). Subjects attaching greater importance to the risk of toxicity than to the likelihood benefit, were classified as being risk averse. Black subjects assigned the greatest importance to the theoretical risk of cancer, whereas White subjects were most concerned with the likelihood of remission. 52% of Black subjects were found to be risk averse compared with 12% of the White subjects ($P < 0.0001$) (19). These results suggest that efforts to improve patient education and physician communication should be made to ensure that all patients have an accurate understanding of the benefits, as well as risks, associated with the best available treatment options.

Liver Disease

Two very different treatment options have been proven to be effective in preventing bleeding in patients with liver disease: medications (beta-blockers) and endoscopic variceal ligation (EVL). Meta-analyses show that, compared with no treatment or a placebo, beta-blockers reduce the risk of bleeding from approximately 30% to 14% over two years. Two recent meta-analyses of studies comparing beta-blockers and EVL show that EVL is marginally more effective than beta-blockers in preventing bleeding without any differences in mortality. Guidelines recommend that beta-blockers be used as first line therapy. In this study we used ACA to elicit patient and physician preferences for both of these options. In direct contrast to current clinical practice and guidelines, we found that 64% of patients and 57% of physicians had stronger predicted preferences for EVL over beta-blockers (20). Despite these trends, our results also demonstrated that predicted preferences vary significantly, indicating that this choice is value-based and emphasizing the importance of incorporating individual patient values into the treatment planning process.

Modified ACA

Experience in conducting ACA surveys has demonstrated that many patients have difficulty understanding the 'importance' questions. Problems arise from (i) the difficulty patients have in incorporating the range in levels into their ratings; (ii) the difficulty of making relative judgments for each attribute, especially if respondents are not able to see all the attributes on a single screen; and (iii) as with other rating tasks, the tendency to overuse the extremes of the scale (21). One option to deal with this difficulty is to simply delete the set of importance questions from the ACA survey. Therefore, this approach cannot be used for investigators requiring output in real time.

In order to address this issue, we developed, in collaboration with Sawtooth Software®, a modified version of ACA importance questions and tested the performance of these questions in a pilot study of patients with knee pain (21). Two questions were developed by the author and programmers at Sawtooth Software® and were subsequently modified based on repeated cycles

of feedback from providers and patients with varied levels of education. In the first question, subjects are presented with a list of all the attributes included in the survey and asked to choose the one that is most important to them. In the second, subjects are asked to rate the importance of the remaining attributes on a grid relative to the one chosen as most important using a numeric rating scale ranging from 'Not nearly as important' to 'Just as important'.

Eligible subjects were recruited at the time of their regularly scheduled initial or follow-up appointment and were randomized to complete the original or the modified ACA importance questions. Both versions were otherwise identical. We included six attributes in the ACA survey: improvement in pain, increase in energy, route of administration, risk of stomach upset, risk of bleeding ulcer, and monthly out-of-pocket cost. Each attribute included three levels, all of which had a natural order except route of administration. After completion of the ACA survey, subjects were asked to (i) rank attributes and treatment options using a card sorting task; and (ii) indicate whether the bar graph depicting the relative importance of each attribute generated by ACA should be 'longer', 'shorter', or was 'just right' (21).

Subjects (N=49) felt that bar graphs illustrating the relative importance were more accurate for the modified version of ACA. The proportion of subjects for which the most important attribute chosen on a card-sorting task matched that generated by ACA was greater for the modified than for the original version (48% vs 29%). The proportion of subjects for which the treatment option chosen on a card-sorting task matched that predicted by ACA was also greater for the modified than for the original version (80% vs 75%). Subjects used a greater number of points to rate the importance of attributes on the modified version of ACA (mean±SD= 3.4 ± 0.9) than on the original version (mean±SD= 2.7 ± 1.0). These findings indicate that the modified version of the ACA importance questions appears to perform as well as or better than the original version (21).

MaxDiff

Cancer Screening

In each of the above studies, a research assistant was needed to administer the ACA task. This remained true even after adopting the modified version of the ACA importance questions described in the preceding paragraph. In our experience, a relatively small proportion of patients would be able to complete ACA surveys independently, from home for example. Given the important goal of developing approaches that can be widely and easily disseminated we have recently begun to explore the use of MaxDiff as a decision support tool for patients. In our first study, we tested the ease of use and the acceptability of MaxDiff to elicit patients' preferences for colorectal cancer (CRC) screening tests (22). Patients filled out the MaxDiff survey while waiting for their doctor's appointment in the waiting room. The survey contained 12 attributes:

1. The rare risk of a problem from sedation that would require hospitalization.
2. How good the test is at finding polyps and preventing cancer.
3. The need to collect your stool and spread a small sample on a card on three separate days.
4. The risk of pain or discomfort from the test.
5. The need to get a ride home after the test.

6. The need to take a day off from work or activities for the test.
7. The need to clean the bowel by drinking a prep before the test.
8. The rare risk that the capsule can get stuck and need to be removed with endoscopy or an operation.
9. The need to have a tube in the rectum to have the test done.
10. The need to have a second test (colonoscopy) to remove a polyp.
11. The rare risk of a tear of bleeding from the procedure that would require an operation.
12. The need to swallow a capsule to have the test done.

The attributes were selected based on literature review and the opinions of six physicians and six patients. 92 subjects were interviewed; 84% were male, and their mean (range) age was 65 (49-80). After performing the survey, 95% of patients reported that the program was easy to use; 97% reported that the program helped them to understand the test options; 92% responded that the program helped them to choose a screening test. Importantly, of the 29% who had refused screening at some point, 85% reported that they would be willing to undergo CRC screening with their preferred test (22). Our results indicate that using patient preferences to guide CRC screening testing could likely improve screening rates, assuming patients have access to all available options.

Kidney Transplant

When a patient is offered a kidney for transplantation they have a choice to accept the kidney or to decline the kidney and remain on dialysis. Transplantation leads to better long-term survival, better quality of life, and lower cost. There are, however, two elements of the decision that validate declining the kidney as a reasonable alternative: change in patient characteristics over time, and heterogeneity of donor kidneys. However, patients may defer and choose to wait for a different kidney without losing their place on the transplant list.

We composed a MaxDiff survey to examine how patients and surgeons prioritize relevant factors when deciding to accept or decline an available kidney (23). In this study, we found that, overall kidney quality was the most important to patients, followed by the function of the kidney at time of death, and the proximity of match. The surgeon's opinion and the risk of contracting a disease both play a significant role in patients' decision making. Time on the waiting list was inversely related to the relative importance of the quality of the donor kidney ($r=-0.30$, $p=0.002$), and function of the donor kidney (-0.31, $p=0.002$). Surgeons were most concerned with overall kidney quality and baseline function, how difficult it is for the patient to find a match (i.e. whether or not the patient is sensitized), and the age of the donor. The patient and surgeon rankings were strikingly similar, sharing five of the first seven factors in common, and seven of the last eight. The notion that doctors and patients seem to consider the same types of factors important is encouraging because it indicates that there is common ground on which to build educational materials such as decision aids to streamline communication (23).

PRACTICAL IMPLICATIONS

These studies have demonstrated that CA/MaxDiff appears to be a very valuable way of quantifying patients' preferences and understanding the impact of specific attributes on patients' choices. The challenge for the health service research community is now to determine how best to 1) develop decision support tools using CA and 2) implement these tools into clinical practice (24).

It is important to note that as patients become more aware of the trade-offs involved, their preferences frequently change. Consequently, depending on how much respondents' opinions evolve, the data generated by the CA task might not be an accurate reflection of patients' newly constructed preferences. In these situations, investigators must decide on the "ideal" amount of education/training to provide patients with before performing the CA task; with greater training being more time consuming and costly, but the more likely to yield accurate preference estimates. "Changing preferences" is less of a concern in situations 1) examining familiar options, in which CA functions more as a tool to elicit, rather than to construct, preferences, and 2) where the investigator is interested in using CA as a vehicle to construct preferences and the outcomes of interest include downstream effects such as patient participation, the quality of informed consent, or patient-physician communication (24).

Although presentation of risk-related information has received the greater part of attention in the literature, how best to present benefits also poses a challenge for CA users. Consider the ways treatment outcomes are often reported. Most pain trials, for example, report mean change in pain or quality of life scales. These data do not easily translate into what patients want to know – which is whether or not the medication will help them – and by how much. Theoretically, both the likelihood and magnitude of benefit should be presented as a single attribute (since both concepts are highly correlated); however, we have found that presenting both concepts simultaneously is difficult for patients to evaluate and overly complicates the task (24).

Another issue, particularly relevant to the US healthcare system, is cost. The wide range of costs associated with different treatment options makes it extremely difficult to examine the influence of this attribute on preferences. Yet, clearly this is an extremely important factor for almost all patients. While it is possible to create different versions of a CA survey for insured and uninsured patients, tremendous variability persists even within these two subgroups. Given the impact of the range of levels on the relative importances generated by CA, and the expected interactions between cost and other attributes, further research is needed to determine how best to include out-of-pocket cost in CA decision support tools.

In most clinical settings, incorporating CA decision support tools into clinical practice will not be possible without significant changes. Common barriers include the difficulty of identifying patients at the point of decision making, insufficient time, the need for support, and lack of space. Most decision tools have been developed for situations for which it is relatively easy to pinpoint the time of decision making, such as elective surgery, cancer screening or treatment for cancer. For chronic diseases, implementing CA as a decision support tool is much more difficult unless a consistent marker for a decision point in clinical care exists. In our experience, lack of appropriate space (i.e. sufficiently private and quiet) has eliminated many potential sites as possible settings for implementation projects. Moreover, because space returns the greatest profit when used for clinical assessments, administrators are reluctant to allocate clinical space for supplementary activities.

Widespread dissemination of CA will likely require development of self-administered tasks (with online and/or telephone support) which can be performed at a time and location convenient for each individual patient. Arguably, however, the best ways to facilitate implementation of CA-based decision support tools in clinical practice would be 1) to lobby for the use of high quality decision support tools to be included as performance measures, and 2) for third party payers to recognize the value of these tools and to reimburse efforts surrounding their use (24).

REFERENCES

1. van Dam HA, van der Horst F, van den Borne B, Ryckman R, Crebolder H**.** Provider-patient interaction in diabetes care: effects on patients' self-care and outcomes. Pt Educ Counsel. 2003;51:17-28.
2. Ward MM, Sundaramurthy S, Lotstein D, Bush TM, Neuwelt CM, Street RJ**.** Participatory patient-physician communication and morbidity in patients with systemic lupus erythematosus. Arthritis Rheum. 2003;49:810-18.
3. Clever SL, Ford DE, Rubenstein LV, Rost KM, Meredith LS, Sherbourne CD, et al. Primary care patients' involvement in decision-making is associated with improvement in depression Med Care. 2006;44:398-405.
4. Fraenkel L, Bogardus S, Concato J, Felson D**.** Preference for disclosure of information among patients with rheumatoid arthritis. Arthritis & Rheumatism. 2001;45(2):136-9.
5. Say R, Murtagh M, Thomson R**.** Patients' preference for involvement in medical decision making: a narrative review. . Patient Educ Counsel. 2006;60:102-14.
6. Montgomery AA, Fahey T**.** How do patients' treatment preferences compare with those of clinicians? Qual Health Care. 2001;10:(suppl.)i39-i43.
7. Suarez-Almazor ME, Conner-Spady B, Kendall CJ, Russell AS, Skeith K**.** Lack of Congruence in the Ratings of Patients' Health Status by Patients and Their Physicians. Med Decis Making. 2001;21:113-21.
8. Beresniak A, Russell A, Haraoui B, Bessette L, Bombardier C, Duru G**.** Advantages and limitations of utility assessment methods in rheumatoid arthritis. . J Rheumatol. 2007;First Release October 15.
9. Lenert L, Kaplan RM**.** Validity and interpretation of preference-based measures of health-related quality of life. Med Care. 2000;38:38-50.
10. Tsevat J, Dawson NV, Wu AW, Lynn J, Soukup JR, Cook EF, et al. Health values of hospitalized patients 80 years or older. HELP Investigators. Hospitalized Elderly Longitudinal Project. JAMA. 1998;279(5):371-5.
11. Winkelmayer WC, Benner JS, Glynn RJ, al. e**.** Assessing health state utilities in elderly patients at cardiovascular risk. Med Decis Making. 2006;26:247-54.
12. Redelmeier DA, Rozin P, Kahneman D**.** Understanding Patients' Decisions: Cognitive and Emotional Perspectives. JAMA. 1993;270:72-6.

13. Kahneman D, Slovic P, Tversky A**.** *Judgement under uncertainty: heuristics and biases.* Cambridge: Cambridge UNiversity Press; 1998.

14. Redelmeier DA, Rozin P, Kahneman D**.** Understanding patients' decisions. Cognitive and emotional perspectives. JAMA. 1993;270:72-6.

15. Fraenkel L, Bodardus S, Wittink DR**.** Understanding patient preferences for the treatment of lupus nephritis with adaptive conjoint analysis. Medical Care. 2001;39(11):1203-16.

16. Fraenkel L, Bogardus ST, Felson DT, Wittink DR**.** Treatment options in knee osteoarthritis: the patient's perspective. Arch Intern Med. 2004;164:1299-304.

17. Fraenkel L, Rabidou N, Wittink DR, Fried T**.** Improving informed decision-making for patients with knee pain. J Rheumatol. 2007;34:1894-8.

18. Constantinescu F, Goucher S, Weinstein A, Fraenkel L**.** Racial disparities in treatment preferences for rheumatoid arthritis. Med Care. 2009;47:350-55.

19. Constantinescu F, Goucher S, Weinstein A, Smith W, Fraenkel L**.** Understanding why rheumatoid arthritis patient treatment preferences differ by race. Arthritis Care Res. 2009;61:413-8.

20. Longacre AV, Imaeda A, Garcia-Tsao G, Fraenkel L**.** A pilot project examining the predicted preferences of patients and physicians in the primary prophylaxis of variceal hemorrhage. Hepatology. 2008;47:169-76.

21. Fraenkel L**.** Feasibility of Using Modified Adaptive Conjoint Analysis Importance Questions. The Patient: Patient-Centered Outcomes Research. 2010;3:209-15.

22. Imaeda A, D. B, L. F**.** What is most important to patients when deciding about colorectal screening? . J Gen Intern Med 2010;25:688-93.

23. Solomon DA, Rabidou N, Kulkarni S, Formica R, Fraenkel L**.** Accepting a donor kidney: an evaluation of patients' and transplant surgeons' priorities. Clin Transplant. 2010;doi: 10.1111/j.1399-0012.2010.01342.x. .

24. Fraenkel L**.** Conjoint analysis at the individual patient level. The Patient: Patient-Centered Outcomes Research. 2008;1:251-3.

PERSONALIZING TREATMENT FOR DEPRESSION: DEVELOPING VALUES MARKERS

MARSHA N. WITTINK[1]
UNIVERSITY OF ROCHESTER MEDICAL CENTER
KNASHAWN MORALES,
AND MARK CARY
UNIVERSITY OF PENNSYLVANIA SCHOOL OF MEDICINE

ABSTRACT

Objective – While "personalized medicine" commonly refers to genetic profiles associated with pharmacological treatment response, tailoring treatments to patient preferences and values is equally important. In this paper, we use the phrase "values markers" to describe a method for creating patient profiles based on the relative importance of attributes of depression treatment.

Methods – Discrete choice analysis was used to assess relative preferences for depression treatment attributes. Preference profiles were developed using latent profile analysis. The subjects were a convenience sample of 86 adults participating in an internet-based discrete choice task. Participants were given 18 discrete choice sets based on type of medication side effect (nausea, dizziness, and sexual dysfunction) and severity (mild, moderate, and severe); and for counseling frequency (once per week or every other week) and provider setting (a mental health, primary care, or spiritual counselor office).

Results – Three profiles were identified: Profile 1 was associated with a preference for counseling and avoidance of medication side effects; profile 2 with avoidance of strong medication side effects and for receiving counseling in medical settings; and profile 3 with a preference for medication over counseling. Persons in profile 1 and profile 2 with more severe depression preferred professional settings over clergy and primary care over mental health settings.

Conclusions – Values markers provide a foundation for personalized medicine and reflect current initiatives emphasizing patient-centered care and may prove useful in tailoring depression treatment to enhance adherence and outcomes.

INTRODUCTION

The notion of "personalized medicine" has been gaining increased attention in the area of mental health.(10,11) Typically this concept refers to the field of pharmacogenomics, which can be deployed clinically to stratify patients into treatment responders and treatment non-responders based on genetic profiling.(27) However, another vision of personalized medicine is related to tailoring to the values of the patient.(32) Attention to preferences for care can have a favorable

[1] Marsha N. Wittink MD MBE, Assistant Professor, Department of Psychiatry, School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, NY 14642, tel: 585-273-3243, fax: 585-273-1082, E-mail: marsha_wittink@URMC.Rochester.edu. Dr. Wittink was supported by a NIMH Mentored Patient-Oriented Research Career Development Award (MH19931) and a NIMH sponsored grant entitled "Developing Methods for Tailoring Depression Treatment to Older Adults" (R34 MH085906). Dr. Morales was supported by a NIMH Mentored Research Scientist Career Development Award (MH073903).

impact on treatment adherence (33) and subsequent clinical and cost-effectiveness outcomes. (4,14) Analogous to *genetic* markers, profiles of genetic variation related to treatment response,(5) this research seeks to identify *values* markers, profiles of values related to attributes of treatment which may also predict treatment adherence. Although antidepressants and psychotherapy have been shown to be effective in treating major depression, non-treatment or under-treatment for depression remains common. Non-adherence results in increased hospitalizations, health care costs,(1) and mortality (23) and is a major public health concern (22). Our line of research seeks to find specific ways to incorporate management strategies for depression tailored to what patients most value about treatment. Understanding the relative importance individuals ascribe to different features of depression treatment will lead to the next steps necessary to determine how to tailor depression treatments.

Prior research assessing what is valued in depression treatment has focused primarily on patient preferences for counseling or medication treatments (13,17,18) and suggests that the majority of people prefer counseling over medication.(7,18,30) However, few patients receive their preference for counseling. In fact, older primary care patients preferring counseling are less likely to receive depression treatment altogether.(35) While a collaborative care model improves access to counseling (74% in collaborative care vs 33% in usual care),(18) many patients still don't adhere to counseling despite a stated preference for counseling. A potential explanation may relate to unmeasured values regarding particular attributes of treatment. For example, among ethnic minority patients spiritually based treatment may be particularly desirable when it comes to depression treatment.(16,40) If counseling is thought to more commensurate with a spiritual model of depression than medication, patients may express a preference for counseling but be less likely to adhere to it if the therapist doesn't share a spiritual framework for depression. Uncovering what is valued about counseling and other treatment modalities may help us understand the treatment decisions patients make.

While prior studies have reported the prevalence of patient preferences for depression treatment (6,7) systematic approaches have not been applied for creating preference profiles, or "values markers". Values markers can tell us why particular treatments may be preferred, or what constellation of valued attributes is most important to patients; thereby suggesting what might be lacking in conventional treatments. Conjoint methods, first developed in mathematical psychology, (26) are intended to "uncover" the underlying preference function of a product in terms of its attributes In the arena of mental health, Dwight-Johnson and colleagues used conjoint analysis to assess features of treatment that low-income Latinos thought would improve its acceptability (12) and Flach and coworkers applied conjoint in the design of an alcohol and cigarette cessation program.(15)

Our study differs from prior work by estimating individual-level values on specific attributes of treatment to create profiles ("values markers"). Our approach is similar to an approach market researchers use to understand the heterogeneity in consumer preferences in order to target their marketing strategies to particular consumer profiles, a notion known as "market segmentation". Profile analysis has the potential to identify patterns of treatment attributes (values markers) that patients value most -- plausible targets for tailoring interventions to improve initiation, adherence and outcomes.

This investigation had two primary goals building on our previous report.(39) First, we introduce a method to focus tailoring strategies by describing how we create values markers

based on individual-level data. Second, we present how values markers derived in response to a scenario of mild depression change for severe depression. Our work moves beyond the realm of simply suggesting that we give patients the treatment they want (e.g. medication or counseling); instead, our approach may help determine how existing treatments might be improved or tailored to better meet patients' needs.

METHODS

Recruitment

For this preliminary investigation, a convenience sample was employed. Participants comprise a panel who have participated in studies of judgment and decision-making on the world wide web (http://www.psych.upenn.edu/~baron/q.htm) and responded to a request to complete a questionnaire about depression treatment preferences. (2) Approximately 1500 panel members have voluntarily participated over the past 10 years in numerous surveys designed to study decision-making processes (2, 3, 20, 21). The panel is roughly representative of the adult U.S. population in terms of income and education but women are overrepresented. (19) For this study, 500 randomly selected members of the panel were sent an e-mail request with a description of the nature of the task and a URL to the discrete choice task, described below**.** Persons who responded within the first two weeks after the request were participants for this study (n = 86). In this sample, 69% were women and the mean age was 41 years. Participants received $3 as a token of appreciation for completing the questionnaire. This study was approved by the Institutional Review Board of the University of Pennsylvania. No identifying information was obtained nor the participants' experience with or knowledge of depression and its treatment.

Discrete choice conjoint studies involve several steps. The first step involves identifying the attributes of treatment that are most salient to patients. Next, the conjoint task (described below) is constructed based on the attributes and levels identified. Then, employing the choice data, relative preference weights for each attribute are calculated that indicate the contribution of each attribute to the choice. The resulting model can be used to estimate the change in choice expected as levels of the attributes are changed. We describe each step below.

Identification of attributes and levels

We convened 3 focus groups to elicit the salient attributes of treatments: two with adults from primary care settings and one with professionals who manage depression (mental health specialists and primary care doctors). Participants were asked to describe their impression of commonly known depression treatments. Patient participants were specifically asked to describe the positive and negative aspects of depression treatments. The professional group was asked to think about features of depression treatments that might act as barriers or facilitators to patient engagement in treatment. Transcripts from the focus groups were analyzed to select the most frequently mentioned depression treatments and attributes (e.g., side effects of medicines, frequency of counseling sessions) to construct the conjoint task. See **Table 1** for the attributes used for the tasks.

Choice task

We constructed a conjoint task and presented respondents with choice tasks in which attributes of medications would have to be played off against attributes of counseling in selecting preferred treatment. Discrete choice conjoint analysis simulates selection of services or products

in competitive contexts by presenting respondents with a set of products (composed of one level from each attribute) and asking which package they prefer. Because medication and counseling share few attributes, we employed an alternative-specific discrete choice design (25) that allowed us to include attributes relevant and specific to medication and counseling (separately) that might be most influential in patient decision making. Prior to each choice set, participants were given definitions for each attribute and level (see **Table 1**).

Table 1. Definitions of attributes and levels provided to participants.

Medication – You take a pill every day for at least 6 months.

Risk of Side Effects (20%)

- *Nausea*. You have an upset stomach and feel the urge to vomit.
- *Dizziness*. If you were to stand up, you might feel unsteady on your feet.
- *Sexual dysfunction*. You experience reduced sexual interest or drive. Men might experience difficulty with achieving or maintaining an erection.

Severity of Side Effect

- *Mild*. You can easily cope with the side effect. The side effect does not interfere with your day to day functioning.
- *Moderate*. You find it difficult to cope with the side effect and you may need additional medication to treat it. The side effect interferes with some of your day to day functioning.
- *Severe*. You find it very difficult to cope with the side effect and you need additional medication to treat it. The side effect interferes with most or all of your day to day functioning.

Counseling – You schedule appointments to talk with a professional about your life, emotions, and depression and learn new ways to cope with and solve problems.

Number of Counseling Sessions

- *Every week.* You attend counseling for 1 hour every week.
- *Every 2 weeks.* You attend counseling for 1 hour every 2 weeks.

Location of Counseling

- *Primary care doctor's office*. You go to your primary care doctor's office for your counseling.
- *The office of a mental health professional*. You go to the office of a mental health professional for your counseling.
- *Office of a spiritual counselor*. You go to the office of a spiritual advisor (priest, pastor, rabbi, imam, etc.) for your counseling.

Participants were asked to express their preference for medication or counseling based on choice sets presented for a "mild depression" scenario and then for a "severe depression"

scenario. The text preceding each choice task was as follows: "You complained of feeling more tired than usual and you just aren't interested in doing things that you normally enjoy. Your physician diagnoses mild depression and recommends treatment to help. She gives you two choices for treatment. Select which choice most closely resembles the type of treatment that you would like." For the second scenario, the text was the same except that the depression was described as "severe depression."

We selected 18 choice sets with differing attributes of medications and counseling. For medication, 9 combinations were possible; namely, 3 levels of severity (mild, moderate, and severe) and 3 side effects (nausea, dizziness, and sexual dysfunction). For counseling, 6 combinations were possible; namely, 2 schedules for frequency of counseling sessions (once per week or every other week) and 3 locations for the sessions (a mental health professional's office, primary care doctor's office or office of a spiritual counselor). Each choice set required that the respondent choose medication or counseling. With 27 possible combinations for medicines and 6 for counseling, there were 54 possible choice sets. While such a design would allow us to assess interactions between attributes, we decided such a design would place an undue burden on respondents. The experimental design routines in version 9.1 of SAS were used to select a fractional factorial design of 18 choice sets from the 54 possible sets that allowed us to estimate the main effects of each factor.(25)

The analysis proceeded in three stages. First, individual-level relative preference weights were estimated. Our logistic model for calculating individual-level relative preference weights produces a fixed, or overall, effect for each variable, and a random deviation from the overall effect for each respondent. The deviation from the overall effect for each individual is called the random effect because the individual effects are assumed to derive from a random distribution. An emerging standard for analyzing discrete choice data is mixed or random-parameters logit.(34) We use a Bayesian-like approach, called empirical Bayes estimation, because of the advantage of allowing for the estimation of individual-level relative preference weights functions.(28) The empirical Bayes method has the advantage of borrowing information from the entire sample in estimating the random effect for each person. We obtained an empirical Bayes estimate for the individual random effect and the corresponding standard errors. We used the SAS GLIMMIX macro which operates by calling PROC MIXED iteratively to estimate the effects. For all analyses, we used an α of 0.05 to assess statistical significance, recognizing that statistical tests are guides to interpretation and inference.

Latent Class / Profile Analysis

Secondly, we clustered individual preference weights using latent profile analysis or latent class cluster analysis.(37) Latent profile analysis is concerned with deriving information about a categorical latent variable from the observed values of continuous manifest variable. In other words, LPA deals with fitting latent profile models to the measured data. The model groups response profiles into clusters representing preferences for treatment. The response profiles are assumed to be derived from a mixture of class-specific normal distributions, where each class has a unique mean preference weight and corresponding variance. The number of clusters is determined using model fit criterion (e.g. BIC) and clinical relevance.

Finally, to examine the association between the mild and severe depression scenarios, we assigned each individual to the profile for which the model indicated the highest probability.

RESULTS

Values markers for mild and severe depression scenarios

The relative preference weights of depression treatment attributes were used to determine common preference profiles (values markers). The profiles representing values markers in the mild depression scenario are shown in the first three columns of **Table 2**. The last three columns depict profiles for the severe depression scenario. Three profiles were found to best fit the data.

Table 2. Profile analysis of individual-level relative preference weights from conjoint analysis given "mild" and "severe" depression scenarios. Favoring the first attribute mentioned in the row is associated with a positive mean relative preference weight for persons with that profile. A negative mean indicates that the second attribute mentioned was favored. Asterisk indicates estimate is statistically different from zero ($p < 0.05$).

	Mild depression			Severe depression		
	profile 1	profile 2	profile 3	profile 1	profile 2	profile 3
Type of treatment						
medication vs. counseling	-5.68*	-0.15	5.86*	-4.79	0.81	6.20
Side effect of medication						
nausea vs. sexual dysfunction	0.17	0.20	-0.27	0.96	1.83	-1.91
dizziness vs. sexual dysfunction	-0.01	-0.65	0.31	0.55	3.01	-1.76
Severity of medication side effect						
mild vs. severe	2.23*	6.82*	-0.93	-1.82	-2.15	3.18*
moderate vs. severe	0.09	-4.52*	2.03*	-0.71	-0.71	1.20
Frequency of counseling						
weekly vs. every 2 weeks	-0.60	0.64	0.31	0.35	0.30	-0.58
Location of counseling						
clergy vs. mental health	0.23	0.95	-0.68	1.76*	-10.35*	1.12
primary care vs. mental health	0.06	0.79	-0.43	-0.23*	2.34*	-0.48
Profile prevalence	**41%**	**18%**	**41%**	**50%**	**13%**	**37%**

Profile analysis under the "mild" depression scenario.

Participants in profile 1 demonstrate a strong relative preference weight for the counseling attribute and for avoidance of the medication side effects attribute. Profile 2 was associated with the avoidance of severe side effects attribute and for the attribute of medical location of counseling. Profile 3 showed a strong relative preference for medication over counseling, and no strong preferences were shown for location of counseling.

Profile analysis under the "severe" depression scenario.

Profiles showed some change in patterns when the depression was severe. Specifically, persons in profile 1 (strong relative preference weight for counseling) and profile 2 (no strong relative preference for counseling vs. medication) preferred professional settings over clergy and for primary care over mental health when the depression was severe. Prevalence of profile 1 (more preference for counseling) increased as well.

Changes in values markers for mild versus severe depression scenarios

The average probability across the sample for classification into profile 1 (relative preference for counseling) was 0.41, for profile 2 (no relative preference for medication or counseling) it was 0.18, and for profile 3 (relative preference for medication) it was 0.41. **Table 3** provides the cross-classification of the profiles for the mild depression scenario (rows) according to the classification for the severe depression scenario (columns). Among 35 persons whose profile indicated a preference for medicine (profile 1) in the mild depression scenario, 11% were classified as preferring counseling (profile 3) when mild was changed to severe depression in the scenario, 20% were in profile 2 (no preference for medication or counseling), and 69% continued to prefer medicines (profile 3). Among 35 persons whose profile indicated a preference for counseling in the mild depression scenario, 5% were classified as preferring medicine when mild was changed to severe depression in the scenario, 11% were in the no preference group, and 83% continued to prefer counseling. For the 16 people who were classified as in the no preference group in the mild depression scenario, 10 (63%) preferred counseling and 6 (37%) preferred medicine when mild was changed to severe depression.

Table 3. Cross-classification of values markers for mild versus severe depression scenarios. Numbers in parentheses are row percents.

		Severe depression scenario			
		Counseling	Intermediate	Medicine	Totals
Mild depression scenario	Counseling	29 (83)	4 (11)	2 (6)	35
	Intermediate	10 (63)	0	6 (37)	16
	Medicine	4 (11)	7 (20)	24 (69)	35
	Totals	43	11	32	86

DISCUSSION

We derived values markers, profiles of preferred attributes of treatment, based on how patients weighed various attributes of depression treatment, identifying three profiles representing a strong relative preference for counseling or medicine, and an intermediate profile associated with location of treatment for mild depression and avoidance of side effects for severe

depression. Among persons classified in the counseling profile, 83% preferred counseling both for mild and severe depression. Among persons who preferred medicine, 69% preferred medicine both for mild and severe depression. The intermediate profile identifies a group whose preferences are changeable and not clearly focused on counseling or medicine. The latter profile may identify a group for whom tailoring an intervention for depression may be most salient.

While accommodating patient values into health care constitutes a key component of overarching principles of health care redesign, precisely how to assess preferences and how to utilize the information in a systematic manner that might facilitate patient-centered treatments remains an area with a need for methodological development. It is important to consider underlying reasons for preferences and to determine whether preferences are related to stable and deeply held beliefs as opposed to relative access to health care information, which may reinforce existing health care disparities.(24)

Before putting our study into the context of current thinking on preferences for depression treatment, several study limitations require comment. First, the participants represented a convenience sample and may react to the hypothetical scenarios very differently than would actual patients. Furthermore, people actually confronting these types of treatment decisions (i.e. patients suffering from depression) might respond very differently to the hypothetical scenarios. Patients who suffer from depression and experience symptoms such as hopelessness, low mood and difficulty making decisions may actually make different choices. With respect to the nature of our conjoint design, side effect severity and type of side effect were only associated with medicine. Had we included attributes related to side effects for counseling we may have seen a different response vis-à-vis the "type of treatment" attribute. In addition, other attributes not included such as cost might have been highly influential. The number and definition of attributes and levels is the critical step in any conjoint analysis task and was based on patient and expert opinion of the most important attributes to include. Our work is novel in that we estimated the individual level relative preference weight for treatment choice and created profiles based on a latent profile analytic model. While others have used individual level- preference weights and cluster analysis to determine preference patterns for treatments,(31) the use of latent profile analysis provides added benefit in that the profiles are model driven and can help determine the presence of an unobserved factor linked to the profile of preferred attributes of treatment.(29) The method we used to calculate individual-level relative preference weights allowed us to learn about groups of persons within the sample with strong preferences for specific attributes. Ascertaining which individuals have strong preferences can direct treatment along lines preferred by the patient, while patients who are in a profile with no strong preference for treatment type (counseling or medication) might need tailored assessment and treatment.

CONCLUSION

Conjoint methods sharpen the focus on "what it is about treatment" that drives preferences and provides specific guideposts for how to design packages of treatments that are patient-centered. Studying how preferences for attributes of treatment are related to treatment adherence, how preferences change over time as depression severity changes and how preferences change with treatment experience are important next steps. In addition, it will be important to look at which types of attribute preferences may represent deeply held beliefs that may be based on cultural norms and traditions, and which types of attribute preferences are more transient and based on limited access to health care information.(24) Conjoint analysis has been successfully

applied to organizational or service redesign to match with changing consumer needs (36, 38) and is increasingly being considered in medical service redesign.(8, 9) For example, conjoint analysis could be used to link patient preferences for specific attributes of both conventional treatments (e.g. medication and/or counseling) and non-conventional depression treatments (such as meditation or spiritual therapy) to observed behavior (initiation and adherence to prescribed treatment). If patients with preferences for specific levels of non-conventional treatment attributes are more likely to be non-adherent to prescribed treatments, then conventional treatments might be adapted to incorporate the desired attributes of non-conventional treatments (e.g. counseling that incorporates aspects of spirituality). Patients who fit into profile 2 identified in this study might be targeted for just such tailoring.

REFERENCES

1. Balkrishnan R, Rajagopalan R, Camacho FT, Huston SA, Murray FT, Anderson RT. Predictors of medication adherence and associated health care costs in an older population with type 2 diabetes mellitus: a longitudinal cohort study. *Clinical Therapeutics.* 2003;25(11):2958-2971.
2. Baron J. *Thinking and Deciding*. 3 ed. New York: Cambridge University Press; 2000.
3. Baron J, Asch DA, Fagerlin A, et al. Effect of assessment method on the discrepancy between judgments of health disorders people have and do not have: a web study. *Med Decis Making.* Sep-Oct 2003;23(5):422-434.
4. Bedi N, Chilvers C, Churchill R, et al. Assessing effectiveness of treatment of depression in primary care. Partially randomised preference trial. *Br J Psychiatry.* 2000;177:312-318.
5. Bourgeron, T & Giros, B. Genetic markers in psychiatric genetics. *Methods in Molecular Medicine.* 2003;77:63-98.
6. Cooper LA, Brown C, Vu HT, et al. Primary care patients' opinions regarding the importance of various aspects of care for depression. *General Hospital Psychiatry.* 2000;22:163-173.
7. Cooper LA, Gonzales JJ, Gallo JJ, et al. The acceptability of treatment for depression among African-American and white primary care patients. *Medical Care.* 2003;41:479-489.
8. Cunningham CE, Buchanan D, Deal K. Modeling patient-centered children's health services using choice-based conjoint hierarchical Bayes. Paper presented at: 10th Annual Sawtooth Software Conference Proceedings, 2004; San Antonio, TX.
9. Cunningham CE, Deal K, Rimas H, et al. Modeling the information preferences of parents of children with mental health problems: A discrete choice conjoint experiment. *Journal of Abnormal Psychology.* 2008;36(36):1123–1138.
10. de Leon J. AmpliChip CYP450 test: personalized medicine has arrived in psychiatry. *Expert Review of Molecular Diagnostics.* 2006;6(3):277-286.

11. de Leon J. Pharmacogenomics: the promise of personalized medicine for CNS disorders. *Neuropsychopharmacology.* 2009;34(1):159-172.

12. Dwight-Johnson M, Lagomasino IT, Aisenberg E, Hay J. Using conjoint analysis to assess depression treatment preferences among low-income latinos. *Psychiatric Services.* 2004;55(8):934-936.

13. Dwight-Johnson M, Sherbourne CD, Liao D, Wells KB. Treatment preferences among depressed primary care patients. *J Gen Intern Med.* 2000;15:527-534.

14. Dwight-Johnson M, Unützer J, Sherbourne C, Tang L, Wells KB. Can quality improvement programs for depression in primary care address patient preferences for treatment? *Medical Care.* 2001;39(9):934-944.

15. Flach SD, Diener A. Eliciting patients' preferences for cigarette and alcohol cessation: An application of conjoint analysis. *Addictive Behaviors.* 2004;29:791-799.

16. Givens J. Ethnicity and preferences for depression treatment. *General Hospital Psychiatry* 2007;29 254-263

17. Givens JL, Houston TK, Van Voorhees BW, Ford DE, Cooper LA. Ethnicity and preferences for depression treatment. *Gen Hosp Psychiatry.* 2007;29:182-191.

18. Gum A, Areán P, Hunkeler E, et al. Depression treatment preferences in older primary care patients *Gerontologist.* 2006;46:14-22.

19. Gurmankin AD, Baron J, Armstrong K. Intended message versus message received in hypothetical physician risk communications: Exploring the gap. *Risk Analysis.* 2004;24:1337-1347.

20. Gurmankin AD, Baron J, Hershey JC, Ubel PA. The role of physicians' recommendations in medical treatment decisions. *Medical Decision Making.* 2002;22:262-271.

21. Gurmankin LA, Baron J. How bad is a 10% chance of losing a toe? Judgments of probabilistic conditions by doctors and laypeople. *Memory & Cognition.* 2005;33(8):1399-1406.

22. Hope C, Wu J, Tu W, Young J, Murray M. Association of medication adherence, knowledge, and skills with emergency department visits by adults 50 years and older with congestive heart failure. *American Journal of Health-System Pharmacy.* 2004;61(19):2043-2049.

23. Irvine J, Baker B, Smith J, et al. Poor adherence to placebo or amiodarone therapy predicts mortality: results from the CAMIAT study. Canadian Amiodarone Myocardial Infarction Arrhythmia Trial. *Psychosomatic Medicine.* 1999;61(4):566-575.

24. Kilbourne AM. Advancing Health Disparities Research within the Health Care system: A Conceptual Framework. *American Journal of Public Health.* 2006;96:2113-2121.

25. Kuhfeld W. *Marketing Research Methods for SAS. Experimental Design, Choice, Conjoint, and Graphical Techniques. SAS 9.1 Edition*. Cary, NC, USA: SAS Institute Inc; 2005.

26. Luce D, Tukey J. Simultaneous conjoint measurement: a new type of fundamental measurement.. *Journal of Mathematical Psychology* 1964;1:1-27

27. Malhotra A, Murphy G, Kennedy J. Pharmacogenetics of Psychotropic Drug Response. *American Journal of Psychiatry.* 2004;161:780-796.

28. Marshall P, Bradlow ET. A unified approach to conjoint analysis models. *Journal of the American Statistical Association.* 2002;97:674-682.

29. McLachlan GJ, Basford. *Mixture Models: Inference and applications to clustering*. New York: Marcel Dekker; 1988.

30. Rokke P, Scogin F. Depression treatment preferences in younger and older adults. *Journal of Clinical Geropsychology.* 1995;1:243-257.

31. Singh J, Cuttler L, Shin M, Silvers JB, Neuhauser D. Medical decision-making and the patient: Understanding preference patterns for growth hormone therapy using conjoint analysis. *Medical Care.* 1998;36(8):AS31-AS45.

32. Stewart M, Brown JB, Weston WW, McWhinney IR, McWilliam CL, Freeman TR. *Patient-Centered Medicine*. Thousand Oaks: Sage Publications; 1995.

33. Thompson J, Scott N. Counseling service features: Elders' preferences and utilization. *Clinical Gerontologist.* 1991;11:39-46.

34. Train K. *Discrete choice methods with simulation*. Cambridge, England: Cambridge University Press; 2003.

35. Unützer J, Katon W, Williams J, Callahan C, Harpole L, Hunkeler E. Improving primary care for depression in late life: The design of a multicenter randomized trial. . *Medical Care.*;39:785-799.

36. Vavra T, Green P, Krieger A. Evaluating EZ-Pass. *Marketing Research.* 1999;11:5-16.

37. Vermunt J, Magidson J. Latent class cluster analysis. In: Hagenaars J, McCutcheon A, eds. *Applied Latent Class Analysis*. Cambridge: Cambridge University Press; 2002:89-106.

38. Wind J, Green P, Shifflet D, Scarbrough M. Courtyard by Marriott: Designing a hotel facility with consumer-based marketing models. *Interfaces.* 1989;19 (January-February):25-47.

39. Wittink MN, Cary MS, Ten Have T, Baron J, Gallo JJ. Towards patient-centered care for depression: Conjoint methods to tailor treatment based on preferences. *The Patient: Patient-centered Outcomes Research.* in press.

40. Wittink MN, Joo JH, Lewis LM, Barg FK. Losing faith and using faith: Older African Americans discuss spirituality, religious activities, and depression. *Journal of General Internal Medicine.* 2009;24(3):402-407.

CONJOINT DESIGN EFFECT ON RESPONDENT ENGAGEMENT THROUGHOUT A SURVEY

PAUL JOHNSON
WESTERN WATS

INTRODUCTION

In 2007, a new conjoint design called Adaptive Choice Based Conjoint (Adaptive CBC) was developed and effectively championed by Sawtooth Software. They showed Adaptive CBC captures more information at the respondent level by identifying non-compensatory rules respondents are using and adapting the questions to respondent preferences (Johnson and Orme, 2007). The extra information provided by Adaptive CBC has also been shown to improve market share predictions when appropriately tuned (Chapman et al, 2009). One aspect of Adaptive CBC not previously examined is respondent fatigue. Johnson and Orme reported that respondents enjoy the adaptive design more even though it takes longer (2007). While increasing the respondent's enjoyment of the survey is vital to preserving respondent's goodwill (Haynes, 2007), it does not measure respondent fatigue. Understanding the effect on respondent fatigue is important, as fatigue can lead to satisficing behaviors like rushing through questions with all the same answer ("straight-lining"), and failing to follow survey instructions (Krosnick et al, 2009).

Respondent fatigue is just one potential cause for question order effects. Schuman et al. found many other causes of question order effects when dealing with surveys about sensitive issues (1981). Question order effects inside of conjoint designs are also well documented for both Adaptive Conjoint Analysis (ACA) (Johnson, 1989) and CBC (Chrzan, 1994). However, placement effects on key survey variables like purchase intent, product satisfaction, and product familiarity placed either before or after the conjoint were not addressed. In this paper we examine both respondent fatigue and question order effects for CBC and Adaptive CBC designs.

METHODS

The study fielded in three waves, each with approximately 800 online panelists from Opinionology's North American panel Opinion Outpost. The first wave, fielded in early March 2010, included two different conjoint designs: Adaptive CBC and a 20 task CBC design. The second wave, fielded in late April 2010, used two variations of Adaptive CBC (combined in this study), and two different lengths of CBC designs (20 task and a 30 task design). This alteration, made at the suggestion of Bryan Orme from Sawtooth Software, accounts for the additional time required to complete an Adaptive CBC design. The third and final wave fielded in late July/August 2010 and used an identical design as the second wave. In all three waves respondents were randomly assigned to one of the designs and then randomly assigned to see the conjoint design either before or after the set of key response variables. Figure 1 depicts the overall experimental design across all three waves.

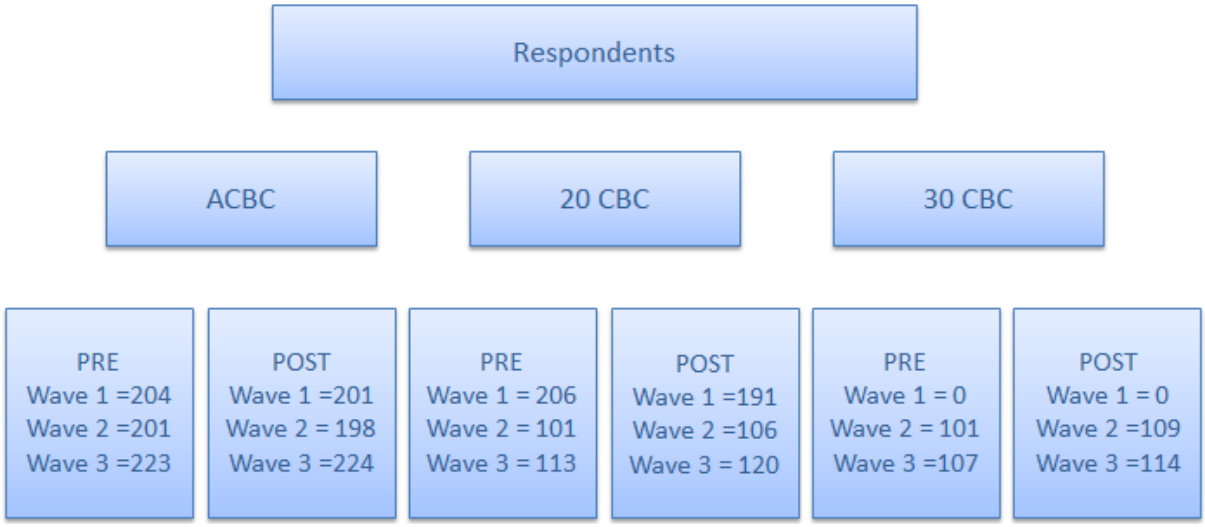


Figure 1.
Sample size and experimental design of the study

All three waves used the same questionnaire, the only changes were the type of conjoint and position of the conjoint in the survey. The questionnaire examined attitudes and preferences in selecting a new computer at work. All respondents were asked questions about the respondent's industry, computer purchasing power, and frequency with which they used a computer for various tasks. Those that did not use a computer at work were screened from the sample. Those that used a computer at work, but had no purchasing power over the computer (33%) were monitored across the cells for consistency, but allowed to complete the study. Then the respondents received the experimental combination they were assigned. Lastly, all respondents saw demographic questions which were examined for inconsistency between test designs. No significant differences were found between the cells in the demographics or baseline computer questions, so no weighting was required to account for these differences.

All conjoint designs had a design space consisting of 12 attributes with 2-4 levels inside of each attribute. The conjoint attribute importance ratings and utilities are not directly comparable because the CBC used discrete levels of price while the Adaptive CBC designs used summed pricing, with price as a continuous variable. All designs used the HB algorithm to develop individual level utilities to use in simulation. Hit rate and mean average error (MAE) were used to compare the accuracy of predicting hold out tasks. All respondents saw four hold out tasks immediately after their designated conjoint exercise. The first three tasks were typical CBC hold out tasks with minimal overlap. The last task (in this paper called a "winner task") took the items selected in the three previous tasks and had the respondents select the best out of the three previous selections. Hit rates were then taken by comparing the product with the highest utility to the product selected in the conjoint. The MAE compared the share of preference models after tuning each model in each design.

The key response variables were all tested using a Chi-square overall test for categorical variables or an F-test for numeric variables. Pairwise comparisons were also conducted using the appropriate two sample test. Fatigue metrics such as straightlining, speeding, and following

simple instructions were compared to see how the conjoint design impacted the remainder of the survey.

Conjoint Differences

We first looked at the differences in consistency and accuracy metrics for the conjoint results. For consistency, we used the average respondent root likelihood (RLH) from each design. The results are shown in Figure 2. While the Adaptive CBC and the CBC designs are not directly comparable because of inconsistencies in the number of choices and alternatives, the respondent who saw the conjoint first can be compared to those who saw the conjoint second in each design. The 30 screen CBC design had more consistency when it was shown first, but all other designs did not improve consistency by the placement of the conjoint.

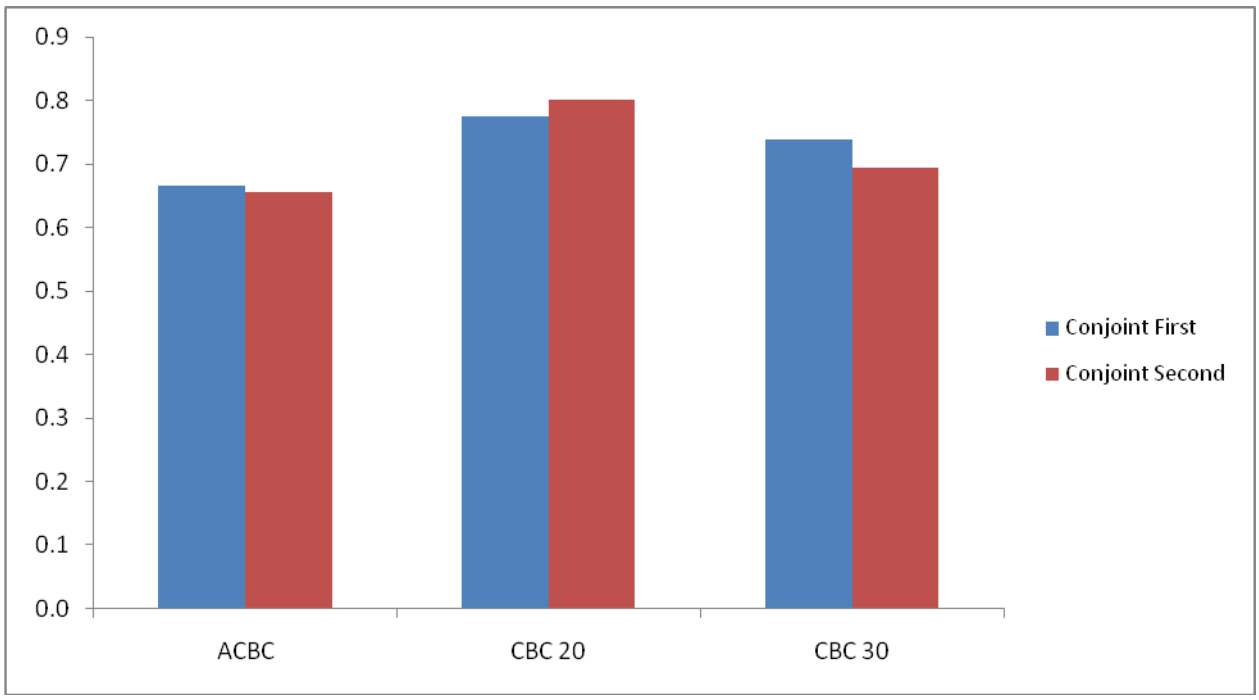


Figure 2.
Average root likelihood (RHL) by design choice and placement

The hit rates are comparable across all design choices and placements. Figures 3 shows the hit rate for the standard CBC holdouts while Figure 4 shows the hit rate for the "winner" holdout. The standard CBC holdouts had a hit rate close to 60% across all test cells while the "winner" holdout had a hit rate close to 50% across all test cells, so overall the prediction accuracy is well over what could be due to chance (33%). While Figure 4 shows directional improvement in the hit rate when the conjoint is seen first, this difference is not statistically significant and could be due to chance.

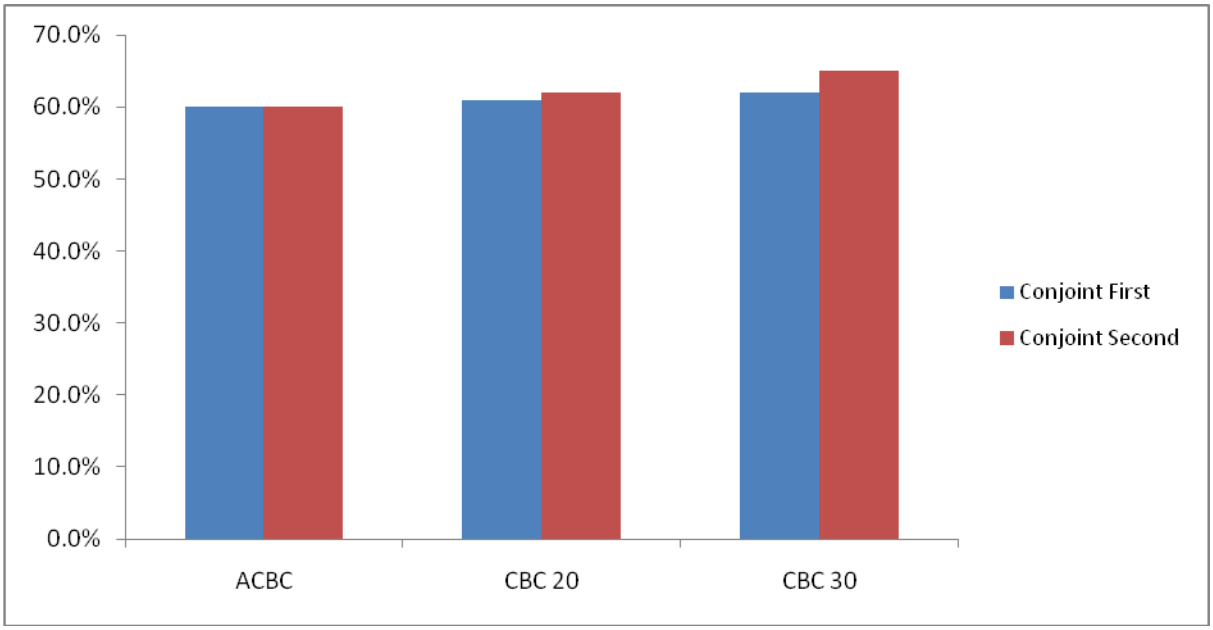


Figure 3.
CBC holdout hit rate by design choice and placement.

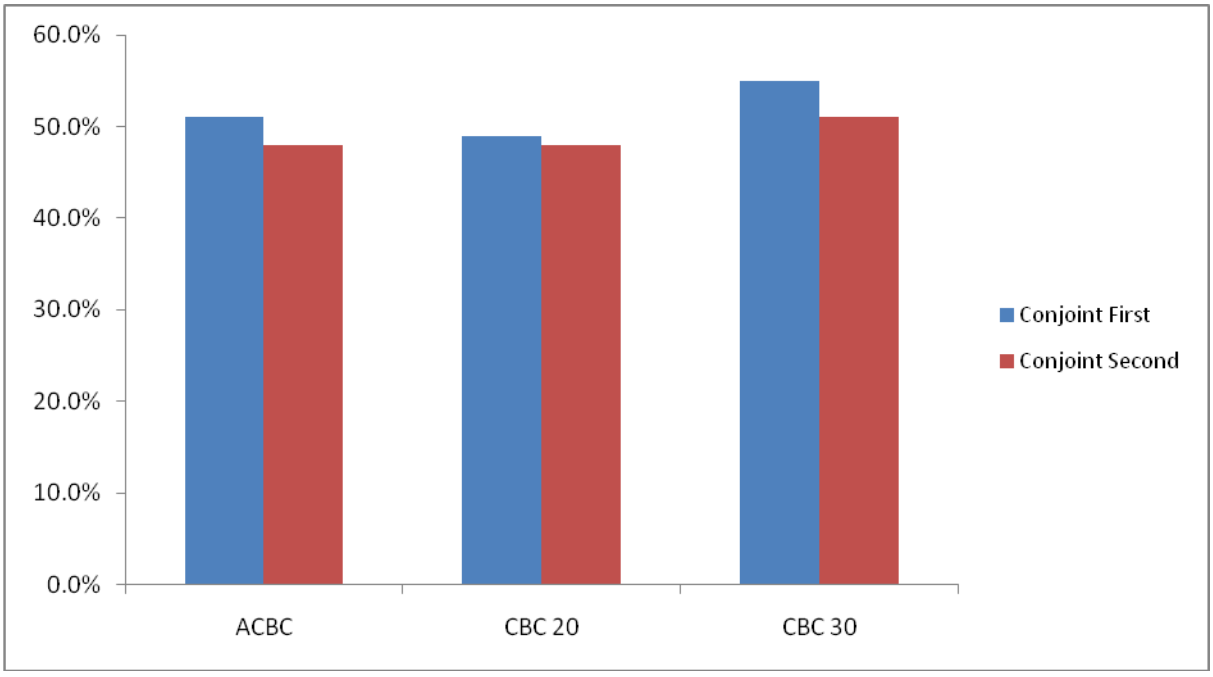


Figure 4.
"Winner" holdout hit rate by design choice and placement.

The MAE was small across all test cell combinations (<5%), but there were differences across the designs and the placement of the conjoint. Figure 5 displays the MAE of the CBC hold outs after tuning. It is not surprising to note that CBC designs performed better on the standard

CBC tasks with no overlap. However, the Adaptive CBC designs do much better on the "winner" holdout tasks. Putting the conjoint first or second doesn't seem to be consistently better across all designs.

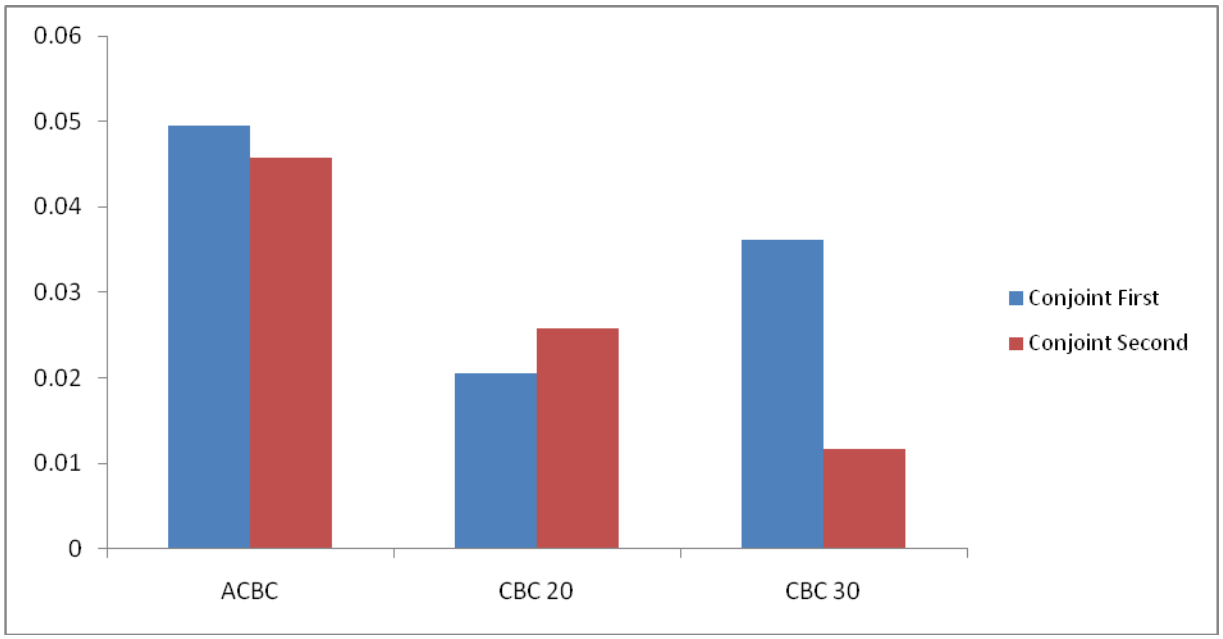


Figure 5.
MAE for CBC holdouts by design choice and placement.

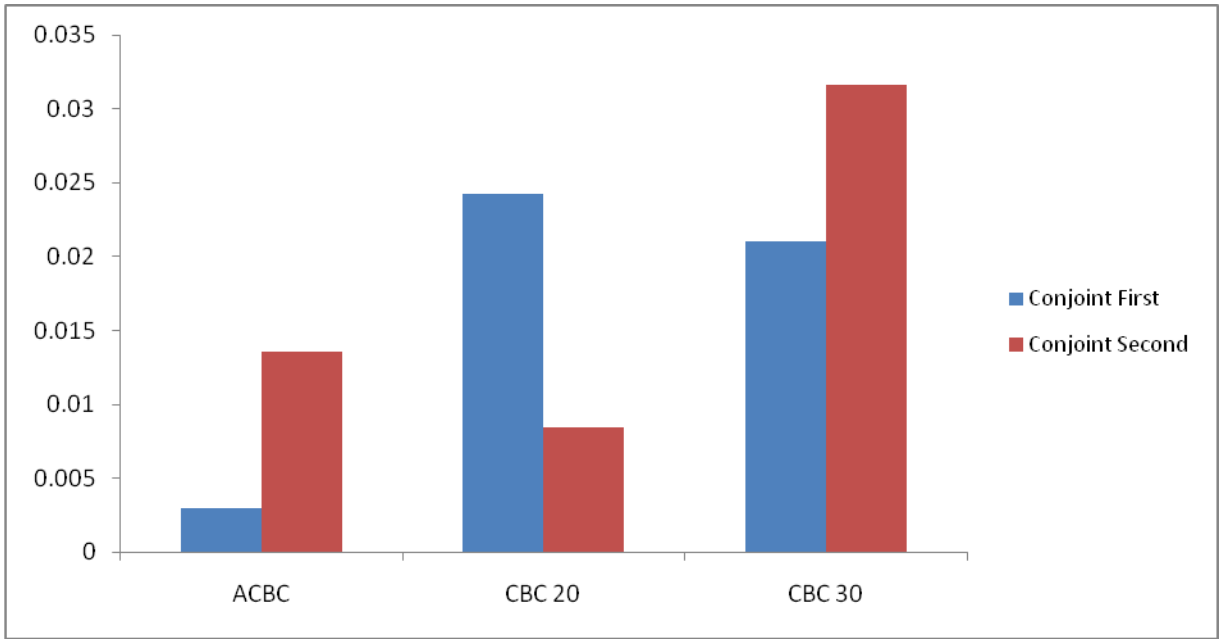


Figure 6.
MAE for "winner" holdout by design choice and placement.

Key Response Variable Differences

We tested four different key metrics in the battery of questions placed either before or after the conjoint design: job satisfaction, product familiarity, product satisfaction, and purchase intent. When testing these variables, we combined all the respondents who saw the conjoint second into one test cell because the conjoint they were assigned does not influence their responses as they have not seen it yet. Figure 7 shows the mean job satisfaction scores which were all measured on a 10 point Likert scale. There is no statistical difference between any of the groups on these job satisfaction questions.

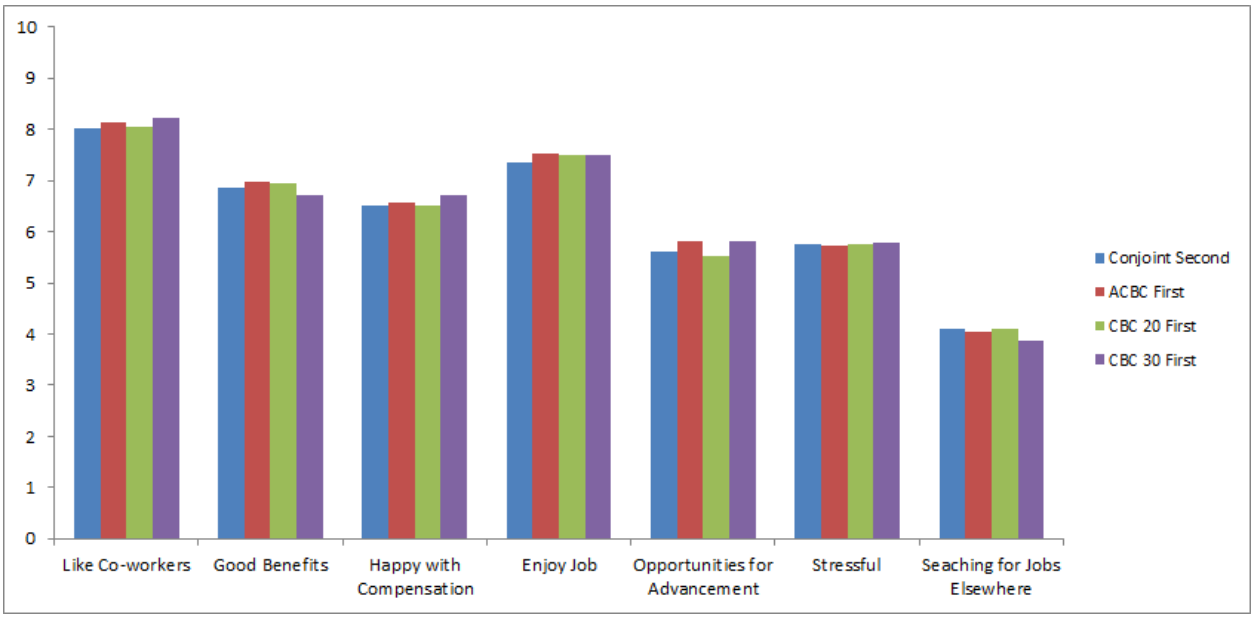


Figure 7.
Mean job satisfaction scores by conjoint design.

We did find a statistically significant, but practically small difference in product familiarity. Any conjoint regardless of design increased the average general familiarity with computers (on a 1-5 ordinal scale) from 4.2 to a 4.4 as shown in Figure 8. This was mostly driven by the percentage of those saying Very Familiar increasing from 42% to 50%. This increase in familiarity did not transfer to how comfortable they are with their current computer, where there is no significant difference. Whether this is because we are asking about comfort rather than familiarity or because we are talking about their specific computer rather than computers in general cannot be determined. Furthermore, whether the magnitude of the learning would be greater with a less familiar product also cannot be determined.

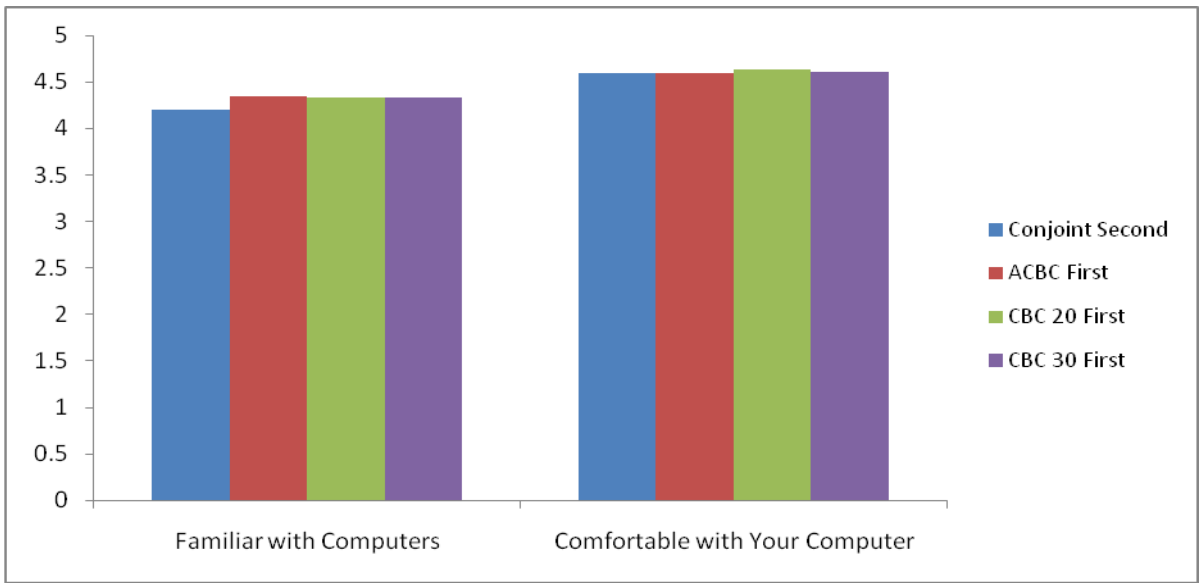


Figure 8.
Mean familiarity ratings by conjoint design.

The satisfaction with their current work computer did not change based on the type or placement of the conjoint design. Figure 9 shows only minor variation in the satisfaction ratings (percent top 2 box) or the percent that have complained about the speed of their current work computer that are not statistically significantly different.

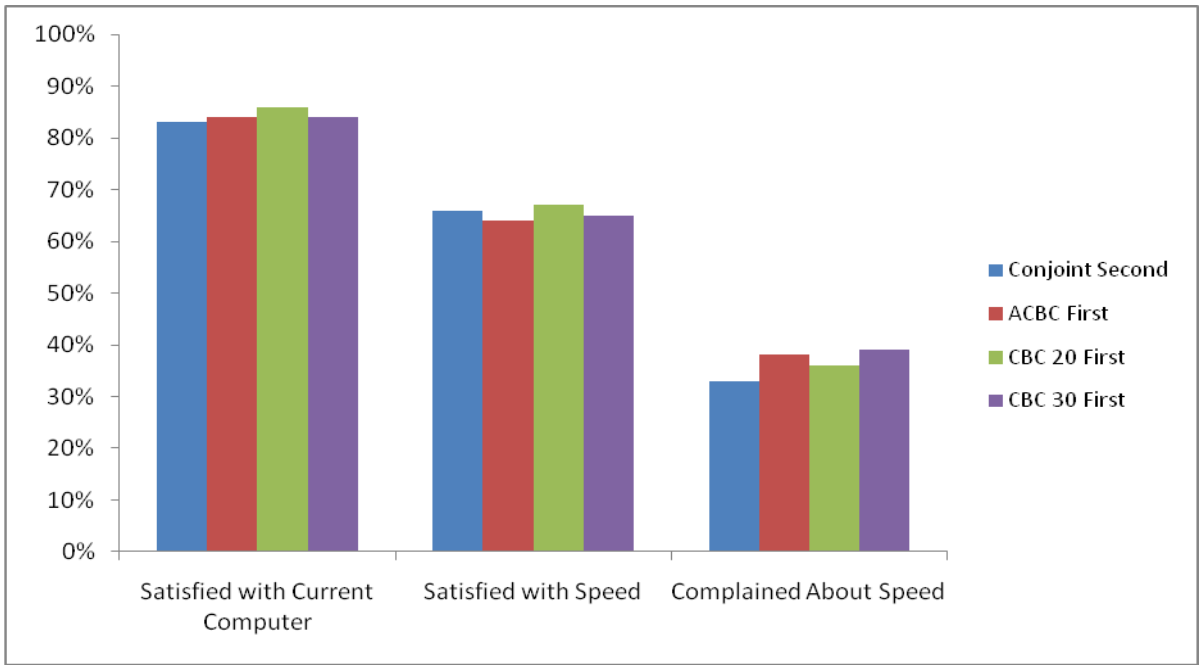


Figure 9.
Top 2 box satisfaction percentages by conjoint design.

We did find an important difference in the purchase intent questions. Those that saw the Adaptive CBC design were significantly more likely to purchase a new computer in the next 6 months than those that saw either standard CBC design. However, none of the three groups were significantly different than those that did not see a conjoint before the purchase intent question. Thus, there is slight evidence that Adaptive CBC designs might increase stated purchase intent in a survey as seen in Figure 10.

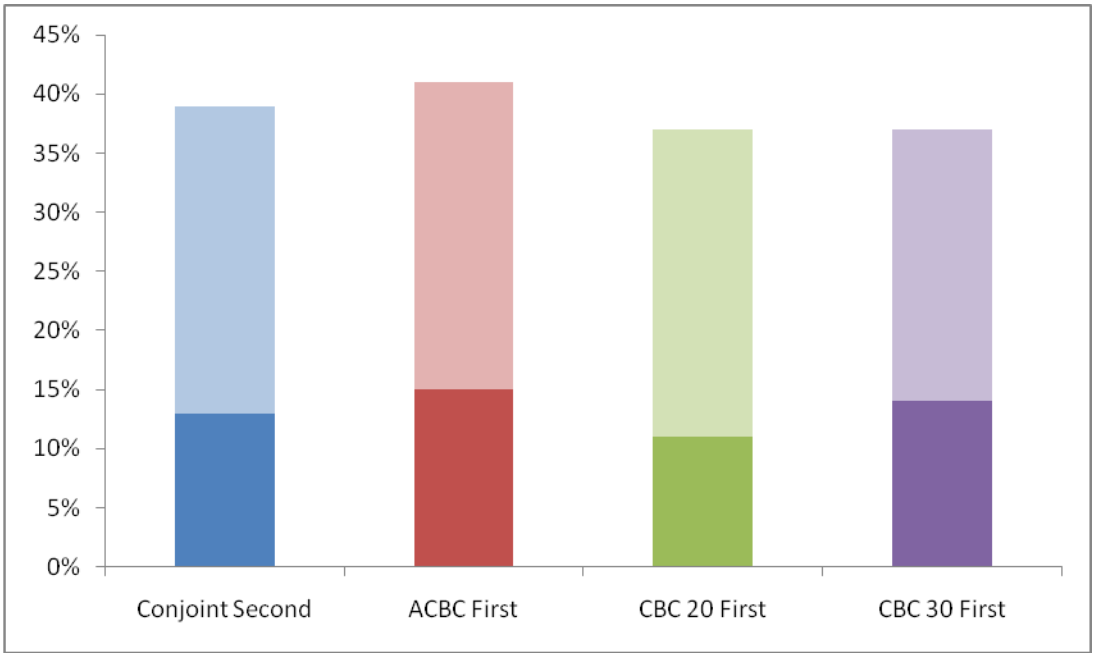


Figure 10.
Top box and Top 2 box purchase intent by conjoint design.

RESPONDENT FATIGUE

We looked at panel fatigue by measuring the length of time that they spent in the survey, the percentage of "straightlining" behavior in the job satisfaction grid, the percentage of selecting non-existent products, and the percentage that fail to follow simple survey instructions. Figure 11 shows all the fatigue metrics with the exception of the time spent in the survey. All the fatigue metrics were fairly low, but the one that caught the most people (across all designs) is the not following simple instructions. In this case, there was a row randomly inserted into the job satisfaction grid that asked specifically for the respondent to select the number three for this row. While overall around 90% of the respondents followed the instructions, there was a significant effect by conjoint design. Those that saw the Adaptive CBC design first were almost 50% more likely (12% as compared to 8%) to not follow the simple instructions in the survey. However, no other respondent fatigue metric showed a significant difference between different conjoint designs.

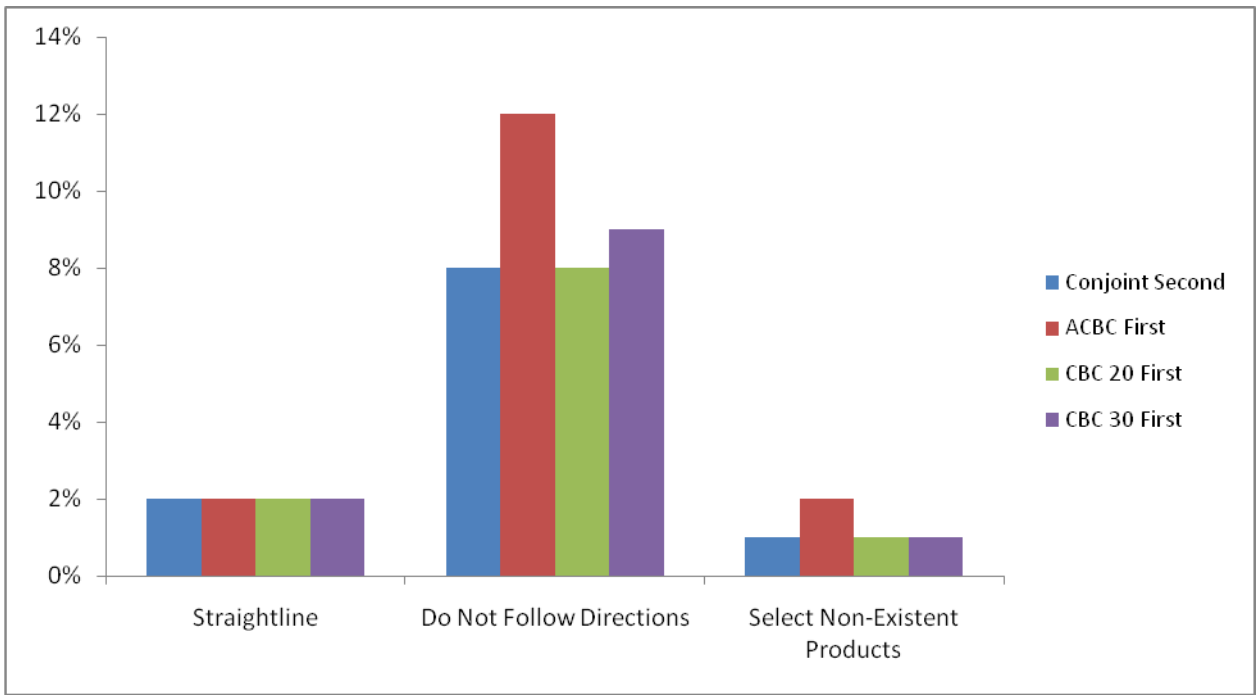


Figure 11.
Respondent fatigue metrics by conjoint design.

While the increase in the propensity to not follow directions indicates that Adaptive CBC fatigues respondents, the time spent in the survey indicates that there is no difference in the fatigue level. If respondents were fatigued after participating in the conjoint exercise, we would expect to see the time spent in the survey non-conjoint section to decrease. Figure 12 shows that time spent in each section of the survey varied by conjoint design and placement. Surprisingly there is no evidence of decreased time in the survey after completing a conjoint exercise. Figure 13 shows that placing the conjoint early in the survey doesn't significantly increase the time spent on the conjoint section. It also shows that the efforts to make a "time equivalent" CBC design were largely unsuccessful. While the 30 screen CBC design did take much longer than the 20 screen CBC design, the Adaptive CBC design still took significantly longer on average. This time difference rather than the questions inside the design itself could potentially have led to the increase in not following simple instructions found later in the survey.

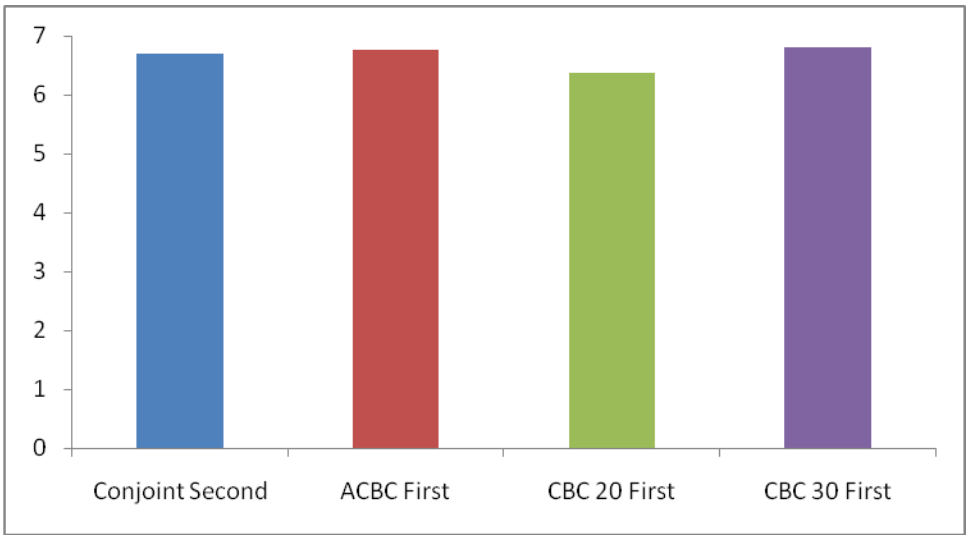


Figure 12.
Average survey completion time excluding the conjoint section (in minutes) by conjoint design and placement.

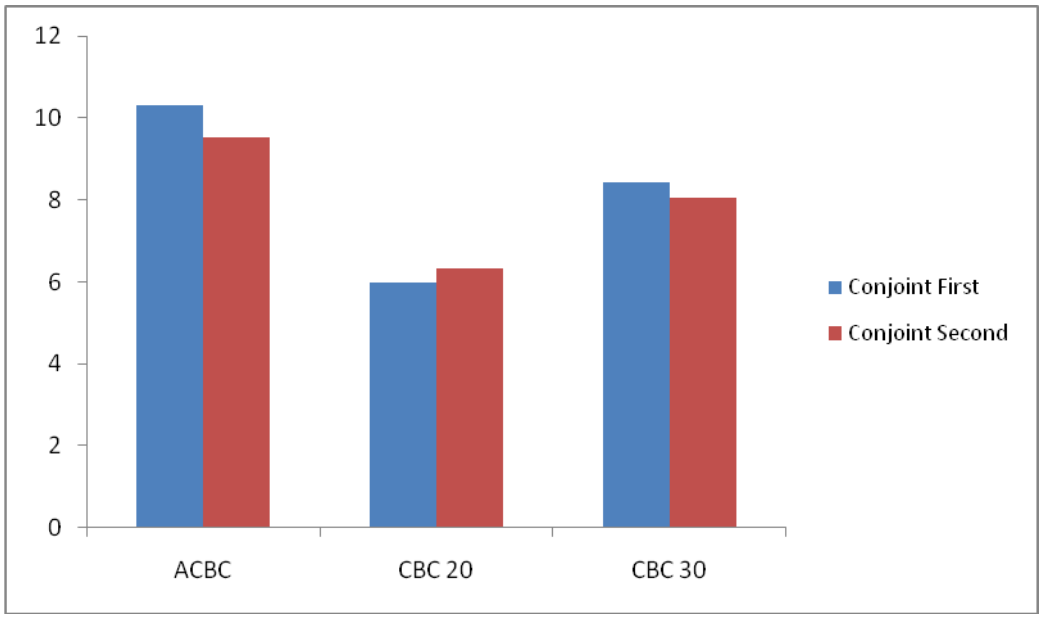


Figure 13.
Average completion time for the conjoint section (in minutes) by conjoint design and placement.

CONCLUSIONS

Because the study was conducted among panel member in Opinionology's North American online access panel Opinion Outpost, the results can only be inferred to this population. The study is also limited to one product (work computers) and the results might not apply to other

less familiar products. Still, the study findings are relevant and can serve as a springboard for verification across other populations and products. The key conclusions from the study are outlined below:

- Respondents who saw an Adaptive CBC design demonstrated signs of respondent fatigue. They were more likely to fail to follow simple directions later in the survey.
- After completing any conjoint design, respondents did not "rush" through the remainder of the survey, but took the same amount of time.
- The Opinionology panelists did not exhibit satisficing behavior typical of high respondent fatigue. Less than 2% selected non-existent products or straightlined.
- In general, the conjoint placement or design did not influence other key response questions in the survey with the following exceptions:
    - Any conjoint exercise increases general product familiarity.
    - Adaptive CBC designs increase purchase intent when compared with standard CBC designs.
- Conjoint placement within a survey has no significant impact on respondent consistency or predictive accuracy at the individual level.
- When looking at aggregate models, Adaptive CBC designs tend to predict "winner" holdouts more accurately.
- Adaptive CBC designs take respondents longer to complete than traditional CBC designs even with 30 tasks.

Works Cited

Chapman, C. N., Alford, J., Lahav, M., Johnson, C., and Weidemann, R. (2009). "CBC vs. ACBC: Comparing Results with Real Product Solutions" *2009 Sawtooth Software Conference Proceedings,* 199-206.

Chrzan, Keith (1994). "Three Kinds of Order Effects in Choice-Based Conjoint Analysis," Marketing Letters 5:2, (1994): 165-172.

Haynes, David (2007). "By the Numbers: Tragedy of the commons revisited." *Quirk's Marketing Research Review* (November 2007), 20.

Johnson, Richard M. (1989). "Assessing the Validity of Conjoint Analysis," *1989 Sawtooth Software Conference Proceedings.* Ketchum: Sawtooth Software.

Johnson, R. M. and Orme, B. (2007). "A New Approach to Adaptive CBC" *2007 Sawtooth Software Conference Proceedings,* 85-109.

Krosnick, Jon., Nie, Norman. and Rivers, Douglas (2009). "Comparing Major Survey Firms in Terms of Survey Satisficing: Telephone and Internet Data Collection" *2009 Annual AAPOR Conference, Miami Beach, FL*

Schuman, Howard, and Stanley Presser (1981). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Content.* New York: Academic Press.

SALES PROMOTION IN CONJOINT ANALYSIS

ELINE VAN DER GAAST AND
MARCO HOOGERBRUGGE,
SKIM

SUMMARY

This paper is about sales promotion as an attribute in conjoint studies. Promotions may involve direct financial gain, and/or indirect benefits. A promotion generates extra attention for the product and the feeling of saving money. Typically, if one does a promotion that has the same financial savings to respondents as lowering the normal price, the effect of the promotion is much higher than simply reducing the price, due to the "attention" effect. It is important to be aware that promotions provide a short-term benefit followed by a post-promotion dip. Even though promotions are difficult to study, conjoint analysis is effective in helping understand which promotion is more effective and which consumers you will attract with the promotion. Future research should aim to incorporate time elements into conjoint studies, to simulate purchase cycles and long-term effects of promotions more accurately.

INTRODUCTION

In times of economic crisis market research is a field that is actually blooming (Andrews, 2008). Especially during times of crisis companies have to make deliberate decisions on how to invest their marketing budget to optimize profits. In the fast moving consumer goods industry, competition is high and promotions are often used as a tool to increase sales. A promotional scheme that will provide the most optimal outcome will give a manufacturer a competitive advantage. Next to boosting short-term sales there are several other motives for using promotions in the consumer goods industry; eliciting trial among non-users or for new product introductions; dealing in markets with increased price sensitivity; and as an alternative for advertising.

In a recent meta-analysis (SKIM, 2009), some of the main promotions found in conjoint studies have been classified as shown in Table 1.

Table 1
Classification of promotion types

Direct Gain	Example	Indirect Gain	Example
Price discount in €	€5 off	Free gift	Free spoon
Price discount in %	20% off	Coupon	Gasoline card €40
Now for	Now €1	Feature	Washable
From - To	From €5.07 to € 4.06	Claims	Best in test
Larger pack	+6 pads		
Additional pack	3 for price of 2		
Multiple unit price	Now 2 for €2		

First of all a distinction is made between a direct and an indirect monetary gain from the promotion, resulting in two different groups of promotions. Direct gain in this case means that the discount is focused on the product itself and the discount comes in either a price reduction or an increase in volume. A gift or a new product feature would be examples of an indirect monetary gain. Within these two groups additional subcategories can be identified. This paper however will focus only on promotions with a direct financial gain.

The effects of a promotion are in general threefold. First of all, promotions are very effective in drawing attention. In addition to this, promotions give the consumer the feeling of having saved money as well as the rational effect of a lower net price. These last two effects are direct benefits for the consumer. When looking at the effects in more detail it is often observed that drawing attention and giving the consumer the feeling of having saved money are in reality more important than the actual net price. This can be illustrated by an example of a study which included promotions. In this particular study there is a base case share of 4.7% at a price of 24.95. First a regular shelf price reduction is applied lowering the price to 19.95. This leads to a share of 5.8%. However, when a promotion ("from 24.95 to 19.95") is used a share of 11.1% is obtained. So basically the rational price effect is only 1.1% (5-8%-4.7%), whereas the effect of drawing attention plus giving the feeling of have saved money is 5.3% (11.1% - 5.8%). This result was also confirmed by a meta-analysis (SKIM, 2009). It must be noted that this effect is not consistently found for all promotion types. It was observed that price promotions always lead to more share, whereas a promotion of i.e. a feature sometimes leads to more share, depending on whether or not the feature is attractive to the consumer.

Another key finding of the meta-analysis on promotions is that the way you express a promotion can lead to different results, even though the actual net price might be the same for the two promotions. This means that one cannot simply replace one promotion by another and expect the effect on share to be the same.

Limitations of Promotions in Conjoint Studies

When using promotions in a conjoint study it is important to keep the limitations in mind. Knowing what can and cannot be determined from including promotions is essential. Only looking at the results as just described in the introduction could lead to over-estimating the effect of a promotion. One of the main reasons for this is that when using current approaches one can only see the direct effect of a promotion. The next section will illustrate why it is so important to determine the full effect of promotion.

Post-promotion dip

In general, most promotions result in an increase in short term sales (Lilien et al., 1992). Using Conjoint Analysis (CA) the direct effect of a promotion can be estimated. However, next to the direct effect of a promotion there is also a dynamic effect. Many researchers have investigated the dynamic effects of promotions. In general it is observed that a promotion period with increased sales is followed by a post-promotion dip. For example, Neslin and Stone (1996) pointed out that stockpiling should be observable in sales data through post promotion dips however it is unclear why these dips are not consistently found in the data. This resulted in a debate in the literature as to why the post-promotion dip was sometimes not observed in store data. In a study by Van Heerde et al. (2000) this so-called paradox of the post-promotion dip was examined and they concluded that the post-promotion dip paradox does not necessarily have to

exist but that it can be noticed in carefully specified time-series models. This finding was also confirmed by other studies (Pauwels et al., 2002; Van Heerde et al., 2004). To be able to determine the best strategy with respect to promotions for a specific product category as well as to understand the post-promotional dip one must comprehend the different promotional responses with respect to changes in consumer behaviour. Multiple researchers (e.g. Chan et al., 2008; Bell et al., 1999; Gupta, 1988) have decomposed the components of short term effects of price promotions into brand switching, purchase acceleration, stockpiling, and increased consumption. These effects can also be grouped into primary and secondary demand expansion (Bell et al., 2002). Secondary demand expansion refers to an increase in sales due to purchases made by consumers that were not buying the brand before (brand switchers). Van Heerde et al. (2004) refer to this secondary demand as cross-brand effects. Primary demand expansion on the other hand relates to additional purchases made by non-switching consumers engaging in purchase acceleration, stockpiling and increased consumption (Van Heerde and Neslin, 2008). This group can be further classified into primary demand borrowed from future time periods (purchase acceleration and stockpiling) and remaining primary demand (increased consumption). The following section will describe the effects of a promotion on consumer behaviour.

Brand switching

One of the first researchers to decompose promotional responses was Gupta (1988). In his study he found that the main response to a promotion was brand switching, accounting for 84% of the change in volume sold. Brand switching in this context refers to the situation that a consumer buys the promoted brand whereas he/she would usually buy a different brand. Gupta's interpretation was that if a brand gains 100 units during a promotion and 74% of the sales elasticity is attributed to brand switching; other brands in the category are estimated to lose 74 units. Other studies have found similar results (i.e. Chiang, 1991; Bell et al., 1999). However, a more recent study by van Van Heerde et al. (2003) re-evaluated the dataset of Gupta (1988) with a different measure. They transformed the elasticity into unit sales and found that only 33% of the volume change during a promotion was due to brand switching. Chan et al. (2008) also confirmed that brand switching is not the main force for increased sales. Nonetheless, brand switching is a major driver behind the sales increase during a promotion, meaning that under promotion users of other brands start buying the promoted brand. Assuming that brand switchers return to their main brand as soon as the promotion finishes, brand switching could not cause the post-promotion dip, as sales would simply return to their average level. In general, brand switching is the only effect that is observed when using promotions in CA.

Purchase Acceleration

Under purchase acceleration consumers decide to make their purchase before they are actually out of stock and hence this affects their regular purchase rate cycle. A promotion can accelerate a purchase (Blattberg et al., 1981) and this effect is closely related to the suggestion of a time-limit or expiration date of a promotion. When consumers believe the promotion is only a temporary offer, they are more tempted to change their behaviour. A study by Inman and McAlister (1994) demonstrated that when the expiration date of a promotion deal is known by consumers there is an increase in redemptions close to this expiration date. The depth of the promotion discount (or the attractiveness of a promotion) is also related to purchase acceleration, meaning that the higher the depth of a discount the more a consumer is inclined to accelerate his/her purchase. Thus purchase acceleration means that a consumer will make his/her purchase

before getting out of stock, and thus the basic purchase cycle will change, whereby contributing to the post-promotion dip.

Stockpiling

Next to purchase acceleration, a promotion can also induce stockpiling, meaning that consumers will purchase more than their usual quantity (Neslin, 2002). This effect is also known as promotion induced stockpiling and whether it has a positive or negative effect for the manufacturer depends on what consumers will do after the promotion (Ailawadi et al., 2007). One result from stockpiling could be that consumers purchase less in the future at the regular price. Chan et al. (2008) showed that the response to a promotion of brand-loyal consumers consists mainly of stockpiling for future consumption. This could be a point of concern because the loyal consumers of the promoted brand are stocking at the discount price, whereas they would have bought the product at the regular price as well. Thus it might be possible that the overall gain of the promotion is negative. Chan et al. also found that brand switchers do not stockpile at all, hence brand switchers will increase sales during the period of promotion. Loyal consumers, however, will purchase more than their average purchase quantity and with this they will delay their next purchase moment.

Increased consumption

The concept of increased consumption is closely related to stockpiling. Bell et al. (2002) defined flexible consumption as increased consumption triggered by the presence of additional inventory. Whether or not consumers will start consuming more under increased levels of inventory depends on the product category (Bell et al., 1999). The occurrence of flexible consumption is very product category specific. With flexible consumption the time between two purchases remains constant over time, even though more of the product has been bought during the promotion period. This illustrates that consumers consume more and faster when they have more stock at hand. This is not the case with pure stockpiling where consumers keep consuming at a constant rate. Bell et al. (1999) report that categories such as bacon, salted snacks, soft drinks and yogurt exhibit flexible consumption, whereas categories such as bathroom tissues; coffee; detergent and paper towels only showed pure stockpiling effects. One could also argue that the combination of stockpiling and increased consumption could eliminate the post-promotional dip: on the one hand a consumer buys more, but at the same time this consumer will also consume more. Hence after the promotion period he/she will need to buy his/her regular quantities again. When stockpiling does not lead to increased consumption, this can be classified as pure stockpiling.

The fact that one can only observe a partial effect of a promotion in a conjoint study leads to questioning whether one should include promotions at all. Promotions can be useful to include depending on what you want to estimate. They can be extremely useful in determining the appropriate way of expressing a promotion. In addition, they can be used to determine what type of consumers the promotion is attracting (consumers from competitors' brands, consumers from own lower tiers etc.). The next section will discuss appropriate ways of implementing certain promotion types in a conjoint study.

Implementation

In conjoint studies we try to mimic reality in the best possible way. Per country and per category the types of promotions that (may) occur are different, as well as the frequency in which

products are on promotion. If the frequency of promotions is very low, it may be recommendable not to include promotions at all. These real life observations about type and frequency of promotions have to be used in order to define the attributes and levels, as well as the experimental design in which we manage the frequency of occurrence of attribute levels. The attribute level "no promotion" (implemented as a blank line) should occur much more often than any of the other promotion levels. Sometimes clients even request to manage the co-occurrence of promotions and SKUs, in order to ensure that all SKUs of a brand are on promotion simultaneously ("line promotion"). Of course a request like this has its implications: it limits the analysis to line promotion scenarios as well. Simulating a scenario with only one SKU on promotion (while this was never shown in a choice task) is no longer valid in this situation. Still, if a client is not interested in single-SKU promotions and only interested in line promotions, this is the way to go.

Generally speaking, there are three different promotion types that are often used in conjoint studies.

1. Gross price shown, net price not shown (merely implied). Examples of this type of promotion are "Now for half price!" or "3 for the price of 2!"

2. Gross price not shown, only net price shown. The typical example is "Now for 9.99!"

3. Gross price shown and net price shown. The typical example is "From 4.29 to 2.99" or merely showing 4.29 with a cross through it and the 2.99 shown beneath it.

The three different promotion types will be discuss briefly on how they translate into attributes and levels in a conjoint design. There may even be a mix of different promotion types within one single study, hence it is also necessary to have a mix of different ways of how to include them as attributes and levels in the conjoint design.

Promotion type A: gross price shown, net price not shown

For this group the original (gross) price is shown to respondents. The net promotion price is not explicitly shown, however this can be implied from the % off or the $ off. To incorporate this type of promotion in the design, two methods are considered. The first method uses a single promotion attribute which is independent from price. This is illustrated below;

Attribute 1

- $3.83
- $4.03
- $4.23
- $4.43
- $4.63

Attribute 2

- (blank = no promotion)
- 3 for the price of 2
- 50% off
- $ 1 off

Basically, the first attribute has a price that reflects the un-promoted price. The second attribute is the promotion attribute. The first level of this attribute is “No promotion,” the remaining levels are the other promotions that were tested in the study.

A second approach would be to use an alternative specific design. This would be done by adding an extra yes/no promotion attribute.

Attribute 1

- $3.83
- $4.03
- $4.23
- $4.43
- $4.63

Attribute 2

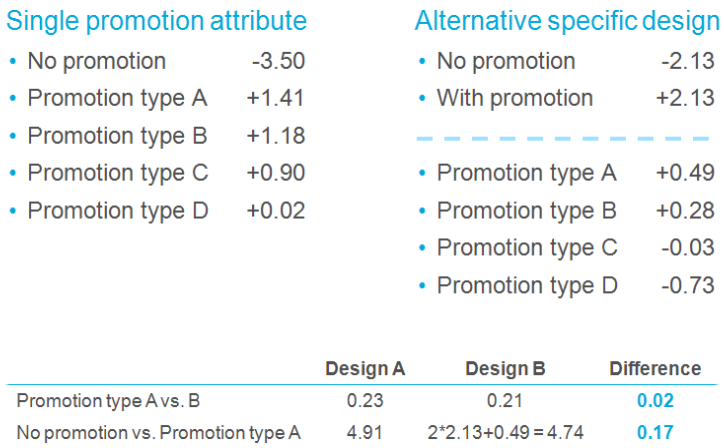
- (blank = no promotion)
- (blank = with promotion)

Attribute 3 (only if attribute 2 =level 2)

- 3 for the price of 2
- 50% off
- $ 1 off

Important to know is whether or not it makes a difference which design to use. If there would be a difference, we would need to know which one is the best – or even the true – way of modelling. The following example compares the two approaches (a single promotion attribute versus an alternative specific design). The mere display of utility values in Figure 1 suggests that there are enormous differences between the two approaches, since the utilities as such do not match with each other at all. However, utilities in themselves are not meaningful; only the difference between pairs of utilities is of real interest.

Figure 1
Comparing single promotion attribute versus an alternative specific design



	Design A	Design B	Difference
Promotion type A vs. B	0.23	0.21	**0.02**
No promotion vs. Promotion type A	4.91	2*2.13+0.49 = 4.74	**0.17**

For example, when taking the difference between "no promotion" versus "promotion type A" it appears that in both designs the outcome is very similar. This is a phenomenon that is in fact very common and it can also be explained very well. In fact the two models are nearly identical because the number of degrees of freedom is the same for both models and every attribute level in one model can be described as a linear combination of attribute levels from the other model.

There is only one substantial difference between the two models. This difference has not been shown explicitly up till now and in practice the difference appears to be very minor. Namely, the constraints on the utility values (may) work out slightly different. In the alternative specific design we can only constrain the utility of "no promotion" as to not to exceed the utility of "with promotion." Consequently the utility of "no promotion" does not exceed the *average* of the promotion utilities. In the other model we can have constraints on the utilities of *each* promotion utility versus the "no promotion" utility. So this is stricter. But as mentioned earlier, in practice it hardly makes a difference.

Promotion type B: gross price not shown, net price shown

The second type of promotion that is investigated only shows the net promoted price. "Now For" is the typical representative of this type of promotion. Since gross price is not being shown while there is promotion, it cannot be regarded as an independent attribute from promotion. So we need another approach. In fact there are two approaches possible which are interchangeable. The first approach uses a single attribute for promotion and price together. This can be done because price and promotion are in this context mutually exclusive alternatives.

Attribute 1

- $3.83
- $4.03
- $4.23
- $4.43
- $4.63

- Now for $2.49 !
- Now for $1.99 !

The other approach is again making use of an alternative specific design. This may look as follows:

Attribute 1

- Without promotion
- With promotion

Attribute 2 (only if attribute 1=1)

- $3.83
- $4.03
- $4.23
- $4.43
- $4.63

Attribute 3 (only if attribute 1=2)

- Now for $2.49 !
- Now for $1.99 !

Although these two approaches look different the models are essentially identical. The display of utility values seems different however the simulation results are the same for the two models. This is because the number of degrees of freedom is the same for both models and every attribute level in one model can be described as a linear combination of attribute levels from the other model.

Promotion type C: gross price and net price shown

The last promotion type that will be discussed in this paper is "From-To." The difference between the original and promoted price is meant to make the promotion offer extra attractive. With this promotion the original price is being shown to respondents. In addition to this the promotion price is also shown (and often highlighted in a task to create some more attention). To incorporate this type of promotion in your design there are several possibilities. The easiest method is to ignore the original price and treat the promotion the same as "Now For." However, a drawback of this method is that ignoring the original price might lead to biased results. Since the "From" price might change from one scenario to the other the change in consumer behaviour could potentially not be reflected correctly.

A second more extensive model could be to keep the original price as a main effect. The original and the net promoted price would be used as independent attributes. At some point, simulation results would become illogical though: suppose the gross price decreases and the net price remains constant. When constraints are correctly applied, the share of preference would increase in this scenario. However, we would rather expect the opposite because the gap becomes smaller, hence the promotion becomes less attractive; or we would expect nothing to change because the net price remains the same.

The third option is the most extensive, i.e. with the largest number of degrees of freedom. It is namely keeping original and net promoted price, and also including all interactions between original and promoted prices in the model. In theory this could solve the problem of the second model. But the effect is in fact quite uncertain because this model can be steered less with constraints on utility values (the interaction effects are unconstrained) and the number of degrees of freedom may be large.

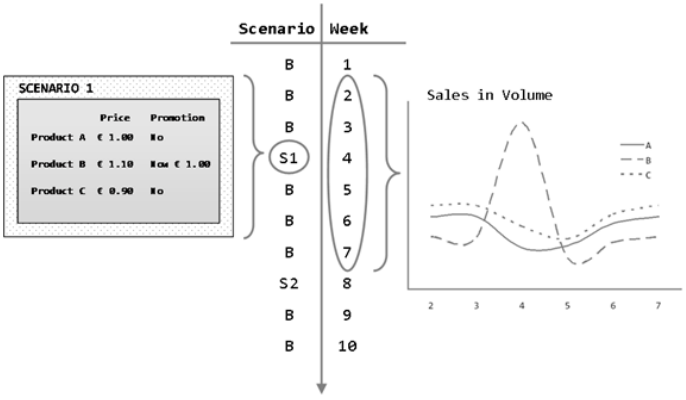
The latter remark implies that a choice has to be made when designing the study: either the number of net promoted prices (per SKU) should be kept at a minimum in order to accurately estimate the interaction utilities or the simple main effects model should be used, thus ignoring the effect of the gross "From" price.

Future Developments

The current approach of using promotions in conjoint studies is still subject to many limitations. The fact that one cannot see the long term effect makes it not possible to determine the eventual net worth of doing a promotion. In addition it is also not possible to optimize a promotional scheme over a longer period of time in terms of length of a promotion as well as the interval between promotions.

In order to deliver more insightful results for clients we should move away from our current static approach of simulating to a more dynamic approach. Rather than just seeing the direct effect of a promotion, as is the case with the static approach, one would like to see the effect on several weeks. One way to make the approach more dynamic would be to simulate individual purchase cycles. Using these individual purchase cycles one would be able to determine for each period of time (say a week) the choices each individual would make. The most important choices to determine are; when the purchase will be made, what product will be bought, in what quantity and how much stock this individual will have. In order to simulate this one will need to have more information regarding the consumer's behaviour, which can for example be obtained from diagnostic questions. Using information of normal purchase behaviour such as purchase frequency, purchase quantity and consumption levels the individual purchase cycle can easily be simulated using this information in combination with the results from the conjoint. In addition, historical sales data will be required to estimate a part of the promotion effect. One might argue that the same results could be obtained with just doing a time series analysis on historical data, however the combination with the conjoint data will give more detailed results on a SKU level. To illustrate the concept a little bit better refer to Figure 2.

Figure 2
Dynamic approach



In Figure 2 one can see a timeline of 10 weeks—weeks here being the unit of time. For each week a specific scenario can be applied. After defining the timeline the purchase cycle for each respondent will be estimated and the aggregate of all sales will represent the volume share. The volume share will be calculated per unit of time; however it can also be estimated as an average of a longer period of time in order to get a better sense of the overall effect.

Unfortunately, more insight comes at the expense of a more elaborate model and increased complexity in the analysis phase. The new approach might in addition still be subjected to assumptions made by the researcher.

Conclusions

Promotions are a very popular marketing tool and therefore are often included in conjoint studies. Promotions can be classified into two groups; promotions with a direct financial gain and promotions with indirect benefit. The first group is most often implemented in conjoint studies. The implementation section showed how certain types of promotions should be implemented in a conjoint study as well as the differences between different methods.

Including promotions in conjoint studies still yields some limitations, most importantly the fact that only the direct effect is captured. Moving to a dynamic simulation approach whereby a time element would be included could be one potential idea, especially to cover the phenomenon of "post promotion dips." This would make the results found from promotions in conjoint analysis more useful and realistic to clients.

REFERENCES

Ailawadi, K., Gedenk, K., Lutzky, C., Nelsin, S., 2007. Decomposition of the sales impact of promotion-induced stockpiling. Journal of Marketing Research 44, 450-467.

Andrews, A., May 2008. Outlook bright for market research industry. Timesonline.

Bell, D., Chiang, J., Padmanabhan, V., 1999. The decomposition of promotional response: An empirical generalization. Marketing Science 18, 504-526.

Bell, D., Iyer, G., Padmanabhan, V., 2002. Price competition under stockpiling and flexible consumption. Journal of Marketing Research 39, 292-303.

Blattberg, R., Eppen, G., Liebermann, J., 1981. A theoretical and empirical evaluation of price deals for consumer durables. Journal of Marketing 15, 116-129.

Chan, T., Narasimhan, C., Zhang, Q., 2008. Decomposing promotional effects with a dynamic structural model of flexible consumption. Journal of Marketing Research 45, 487-498.

Chiang, J., 1991. A simultaneous approach to the whether, what and how much to buy questions. Marketing Science 10, 297-315.

Gupta, S., 1988. Impact of sales promotions on when, what, and how much to buy. Journal of Marketing Research 25, 342-355.

Inman, J., McAlister, L., 1994. Do coupon expiration dates affect consumer behaviour? Journal of Marketing Research 29, 423-428.

Lilien, G., Kotler, P., Moorthy, K., 1992. Marketing Models. Prentice-Hall, Englewoon Cliffs, NJ.

Neslin, S., 2002. Sales promotion. Tech. rep., Cambridge, MA: Marketing Science Institute.

Neslin, S., Stone, L. S., 1996. Consumer inventory sensitivity and the postpromotion dip. Marketing Letters 7, 77-94.

Pauwels, K., Hanssens, D., Siddarth, S., 2002. The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. Journal of Marketing Research 39, 421-439.

SKIM, 2009. 5 easy tips for getting the highest return on your money. White paper.

Van Heerde, H., Gupta, S., Wittink, D., 2003. Is 75% of the sales promotion bump due to brand switching? No, only 33% is. Journal of Marketing Research 40, 481-491.

Van Heerde, H., Leeflang, P., Wittink, D., 2000. The estimation of pre- and post-promotion dips with store-level scanner data. Journal of Marketing Research 37, 383-395.

Van Heerde, H., Leeflang, P., Wittink, D., 2004. Decomposing the sales promotion bump with store data. Marketing Science 23, 317-334.

Van Heerde, H., Neslin, S., 2008. Sales Promotion Models. Springer.

How Many Questions Should You Ask in CBC Studies? – Revisited Again

Jane Tang and Andrew Grenville
Vision Critical

Abstract

Does quantity equal quality? Is the perfect the enemy of the good? What role does respondent engagement play in data quality? These are all questions central to the issue of the tradeoff between more choice tasks versus more respondents, in this paper on CBC in the era of on-line panels.

It has long been known that respondents complete tasks later in a survey much more quickly than earlier tasks. This increase in speed has usually been attributed to increasing familiarity with the task. By examining the difference between the earlier and later tasks, we conclude that respondents are also using more heuristic simplifying rules as they proceed through the survey, showing symptoms of being becoming less engaged in the process.

Prior research has shown that keeping surveys short has advantages in the short term (engaged respondents give more accurate answers) and long term (healthier panels overall). We conclude that a shorter CBC task results in only a small loss of model precision. Modeling based on short tasks also has increased sensitivity and better internal consistency.

We would like to thank Michael Mulhern of Mulhern Consulting for our discussion that generated the initial idea of this paper and for his thoughtful comments.

1. Introduction

There are several prior papers that focus on this topic. The first was written in 1996 by Rich Johnson and Bryan Orme. Their analysis was based on 21 data sets contributed by Sawtooth 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 determine that respondents can answer at least 20 choice tasks without degradation in data quality. Later tasks provide data at least as reliable as earlier tasks, and they are often completed much faster. Respondents take about one-third as much time answering the last tasks (12 seconds, versus 35 seconds for the first task). Furthermore, the "brand" factor becomes less important while the "price" factor becomes more important in later tasks. Respondents are more likely to use "None" in later tasks, although the authors find no evidence of this being caused by decision avoidance. Johnson and Orme concluded that you can easily trade-off between sample size and task length (at least for aggregate logit analysis). Doubling the number of tasks per respondent is about as effective in increasing precision as doubling the number of respondents.

Markowitz & Cohen (2001) dealt with the same tradeoff issue, but in a Hierarchical Bayes (HB) framework. They approach this problem using simulated data of various sample sizes and model complexity. They conclude 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.

Hoogerbrugge & van der Wagt (2006) is the second paper with which our title quotes, focusing on holdout board hit rate prediction. They find that 10-15 tasks are generally sufficient for the majority of studies. The increase in hit rates beyond that number is minimal. The complexity of the model has an important effect on the absolute magnitude of the hit rate.

These prior research papers suggest that a choice section of ten to fifteen tasks (up to 20 depending on model complexity) is optimal for HB estimation. Among practitioners, the focus is often on pushing the respondents hard with long choice tasks while keeping the sample sizes small.

Today, the widespread use of Internet panel samples throws some doubts on these results. We see repeated complaints about the length and repetitiveness of choice tasks in the verbatim feedback from our panelists. Even our clients often question the large number of tasks, having become fatigued when trying the survey for themselves. As panel owners and operators, we feel obliged to keep respondent burden to a minimum. Maintaining a panel of engaged respondents is key to us and to our clients.

Two recent papers help us further refine our thinking.

Hauser, Gaskin & Ding (2009) conclude that heuristic non-compensatory rules are less likely when consumers are faced with simple choice decisions; with few alternatives and with low time pressure. The authors also conclude that consumers who are more familiar with the category are more likely to use heuristic rules in their decision making. These results imply that respondents may do the same thing when presented with long choice tasks. That is, if the choice exercise is structured with many choice tasks, each with many alternatives, respondents are more likely to rely heavily on heuristic simplifying rules. We also hypothesize that those who are more familiar with the category may behave differently when it comes to using these rules.

Suresh & Conklin (2010) examine the issue of respondent engagement and its impact on choice modeling. While the choice section shown to three cells of respondents is identical, they are first put through brand attribute sections of varying complexity. The authors conclude that complex survey design leads to lower respondent engagement. Those respondents who receive the more complex brand attribute section also choose “none” more often and have more price order violations in the choice section.

We are thus motivated to take another look at the number of tasks within CBC. We want to know how today’s panelists react to longer task length, and understand how much shortening CBC tasks would decrease precision in model predictability. We also examine how respondents are able to complete later tasks so much more quickly. While familiarity is a factor, we suspect that fatigue, and simplifying rules respondents use due to fatigue, might also be at play.

Our study focuses on these issues.

2. Study Design

We use a three-cell experiment as follows:

Cell A (n=1200): 6 choice tasks per respondent, each with 3 alternatives

Cell B (n=500): 15 choice tasks per respondent, each with 3 alternatives

Cell C (n=300): 15 choice tasks per respondent, each with 5 alternatives

The sample sizes are designed so that approximately the same total number of alternatives (approximately 22,000) is shown in each cell.

We have three objectives:

1. We want to know if respondents spend less time on earlier choice tasks when they are faced with a larger number of tasks. We also want to know whether they spend longer on tasks with more options (i.e. more complexity).

2. We want to assess if more choice tasks and more complex choice tasks lead to more precise and better quality models. Model precision is defined by predictability, i.e. the model's ability to predict both holdout tasks and "real world" behaviour. Model quality is defined by sensitivity and consistency as follows:

* Sensitivity: the model's ability to differentiate preferences for factor levels.

* Consistency: the model's ability to provide estimates consistent with known logical preference in the "price" factor.

3. Lastly, we want to determine if respondents are more likely to use simplifying rules in later tasks, and if those respondents who are more familiar with the category behave differently from those who are not familiar with the category. If respondents are indeed simplifying more in later tasks, we expect the relative importance of their most important factors would go up. Their later tasks will also exhibit less utility maximizing and less tradeoff behavior, resulting in less "matching" between their chosen alternatives and their "ideal" option.

Our focus is on using CBC to understand the tradeoffs respondents make when it comes to political policy platforms. The questionnaire consists of the following sections:

- Most important issue facing the US today;
- Performance of the President and Congress: these are used as external validity checks;
- A build-your-own (BYO) section that serves both as the introduction to CBC factors and provides an opportunity for respondents to describe their "ideal" candidate's policy platform;
- Choice task section for each cell: respondents are told how many tasks they will do. No warm up task is used in the choice section;
- Holdout task;

- Additional questions to assess external validity: Voting intention 2010, party affiliation, last vote (2008), political leaning (conservative vs. liberal); and
- Respondent feedback questions.

In Tang & Grenville (2009), we found that a BYO section is very useful in engaging respondents' attention on the factors that are involved in the decision process. It is a commonly used technique that often serves as the introduction to a CBC task. For example, the Sawtooth Adaptive CBC product (ACBC) starts with a BYO task for the respondent to build his "most likely" product.

In consultation with our public affairs colleagues at our sister company, Angus Reid Public Opinion, we choose five areas of political policies within the CBC section, each with two levels describing the traditional Democratic and Republican party positions. A third "no mention" level is added to each area. These areas are chosen because they are considered to be currently important issues.

HEALTH CARE	(R) The government should get out of health care.
	(D) The government should provide health insurance for all Americans.
FOREIGN AFFAIRS	(R) Overseas America should focus on leading the world and promoting our values, and not listen to the UN.
	(D) America should always work with the UN and other countries to improve the international situation.
SIZE OF GOVERNMENT	(R) The federal government is bloated, corrupt and wasteful - spending needs to be cut dramatically.
	(D) The federal government needs to spend more to provide high quality social programs for all Americans.
ENVIRONMENT	(R) Jobs, a strong economy, and energy independence are more important than the environment.
	(D) We need to invest in clean energy to build a green economy to fight climate change.
EDUCATION	(R) The best way to improve the education system is to encourage more charter and independent schools.
	(D) The best way to improve the education system is by giving more resources to public school teachers.

A sixth factor describing federal income tax implication of the policy platform is also added to the design. The tax implication of the policy platform for both a family and a single person is presented to all respondents, as is common in news reporting of this kind.

For a family of four with a household income of $85,000:	**For a single person with an income of $35,000:**
Reduce federal tax by $1,000 per year.	Reduce federal tax by $400 per year.
Reduce federal tax by $500 per year.	Reduce federal tax by $200 per year.
No change to your federal tax.	No change to your federal tax.
Increase federal tax by $500 per year.	Increase federal tax by $200 per year.
Increase federal tax by $1,000 per year.	Increase federal tax by $400 per year.

The holdout task is specially designed to reflect both the Democratic and the Republican parties' traditional positions.

	Candidate A	Candidate B	Candidate C	Candidate D
Health Care	The government should get out of health care	The government should get out of health care	The government should provide health insurance for all Americans	The government should provide health insurance for all Americans
Foreign Affairs		Overseas America should focus on leading the world and promoting our values, and not listen to the UN	America should always work with the UN and other countries to improve the international situation	
Size Of Government	The federal government is bloated, corrupt and wasteful—spending needs to be cut dramatically			The federal government needs to spend more to provide high quality social programs for all Americans
Energy/Environment	Jobs, a strong economy and energy independence are more important that the environment			We need to invest in clean energy to build a green economy to fight climate change
Education		The best way to improve the education system is by giving more resources to our public school teachers.	The best way to improve the education system is to encourage more charter and independent schools.	
For a family of four with a household income of $85,000:	Decrease tax by $500	No change to your current taxes	No change to your current taxes	Increase tax by $500
For a single person with an income of $35,000:	Decrease tax by $200	No change to your current taxes	No change to your current taxes	Increase tax by $200

SAS PROC FACTEX/OPTEX is used to generate block designs for the CBC choice tasks. Respondents from all cells are randomly assigned into one of 15 blocks. Each cell's design is created independently. We use the traditional orthogonal designs, with no attempt at minimum overlap, or utility balancing.

A dual-response approach within the CBC section is utilized. Respondents are asked to choose their preferred candidate, and then are asked how likely it is that they would vote for that candidate. Brazell, etc. (2006) demonstrate that the gains in efficiency in the dual-response approach over the traditional "none" choice approach. This approach has been shown to be valuable when there is a possibility of a large number of "none" choices and preference heterogeneity.

Screen 1 of 6

Which of the following three candidates' policy platform do you prefer the most?

	Candidate A	Candidate B	Candidate C
Foreign Affairs	America should always work with the UN and other countries to improve the international situation.	Overseas America should focus on leading the world and promoting our values, and not listen to the UN.	
Size Of Government	The federal government needs to spend more to provide high quality social programs for all Americans.	The federal government needs to spend more to provide high quality social programs for all Americans.	The federal government needs to spend more to provide high quality social programs for all Americans.
Health Care	The government should get out of health care.	The government should get out of health care.	
Education			The best way to improve the education system is by giving more resources to our public school teachers.
Energy/Environment	Jobs, a strong economy and energy independence are more important than the environment.		Jobs, a strong economy and energy independence are more important than the environment.
For a family of four with a household income of $85,000:	Reduce federal tax by $500 per year.	Increase federal tax by $1,000 per year.	Increase federal tax by $1,000 per year.
For a single person with an income of $35,000:	Reduce federal tax by $200 per year.	Increase federal tax by $400 per year.	Increase federal tax by $400 per year.

- ◯ Candidate A
- ◯ Candidate B
- ◯ Candidate C

Would you vote for the candidate you selected above in a congressional election?

Please select one response only.

- ◯ Definitely would vote for this candidate
- ◯ Probably would vote for this candidate
- ◯ Might or might not vote for this candidate
- ◯ Probably would not vote for this candidate
- ◯ Definitely would not vote for this candidate

Respondents are told how many choice tasks they are expected to do in the introduction page. Within each choice task, a counter "Screen x of 6" or "Screen x of 15" is also displayed so they are aware of their progress.

3. Data & Analysis

The survey was in field from August 13 to 17, 2010. The sample is from Springboard America - Vision Critical's proprietary panel. The sample is balanced to the U.S. population on region, age, gender and ethnicity, and adjusted for response rates. Respondents are randomly assigned into one of the three cells in the study design (proportional to the number of completed interviews required) so response rates and abandonment rates in each cell can be tracked independently.

The panelists respond to approximately 18% of the invites. While the response rates are virtually identical in all three cells, the completion rates in cells B & C are significantly lower than that of cell A. Approximately 25% of the drop-offs in cells B & C appear to come from the additional tasks within the CBC section.

	Cell A (6 tasks, 3 options per task)	Cell B (15 tasks, 3 options per task)	Cell C (15 tasks, 5 options per task)
Sample size (completed surveys)	**1216**	**504**	**308**
Response Rate	**18%**	**18%**	**19%**
Rate of completion	**93%**	**87%**	**89%**

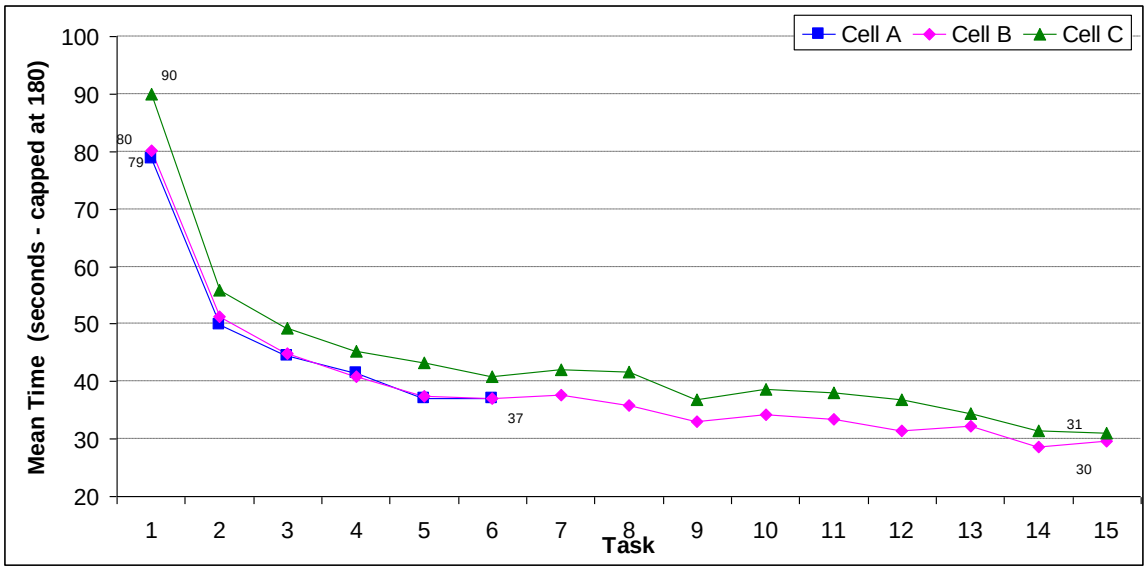
Bold/Red indicates significant difference to cell A, at 5%, two-tailed, ANOVA, Multiple Comparison Bonferroni.
No significant difference between cell B vs. C

We will examine each of our three objectives below.

a. Task Time

We first look at the amount of time respondents spent in each task. We observe the same overall decrease in time spent as respondents complete more and more tasks. Overall, respondents are spending much longer than reported in Johnson & Orme (1996). We suspect this is at least partly due to the effect of using a dual-response format. However, the same one-third rule appears to hold, with task 15 using only approximately one-third of the time of the first task.

Interestingly, cell B respondents (15 choice sets, three options per) were not put off by being told, in advance, that they had 15 tasks to complete. Their pattern of task time is virtually identical to that of cell A (six choice sets, three options per). Cell C respondents (15 choice tasks, five options per) appear to spend more time per task.



A closer look among those who are familiar with politics reveals that while there is an increase in complexity of the choice task in cell C compared to cell B (going from three alternatives per task to five alternatives per task), only those familiar with politics spend more time per task. The additional complexity of the choice tasks has no impact on time spent for those who are less familiar with politics.

Cell	sample size	Total time Spent All Tasks	Time spent first task	Per task Time Task 2-6	Per task Time Task 7-15
Total Sample					
A	**1,214**	**288**	**79**	**48**	**-**
B	**504**	**588**	**80**	**49**	**33**
C	**308**	**656**	**90**	**54**	**37**
Among those familiar with politics					
A	**505**	**292**	**79**	**49**	**-**
B	**201**	**596**	**77**	**49**	**34**
C	**129**	**717**	**96**	**59**	**40**

Green/bold Indicates significant difference to cell B, **plum/bold** Indicates significant difference to both cells A & B, at 5%, two-tailed, ANOVA, Multiple Comparison Bonferroni.
No significant difference between A & B.
No significant difference between the cells among those who are NOT familiar with politics

We define a respondent as being familiar with politics if he is able to give an answer to all 4 of the following questions: Voting intention 2010, Party affiliation, Vote last time (2008), Political leaning (conservative vs. liberal). Respondents familiar with politics make up approximately 40% of the sample in each cell.

b. Hit Rate, Model Sensitivity, and Consistency

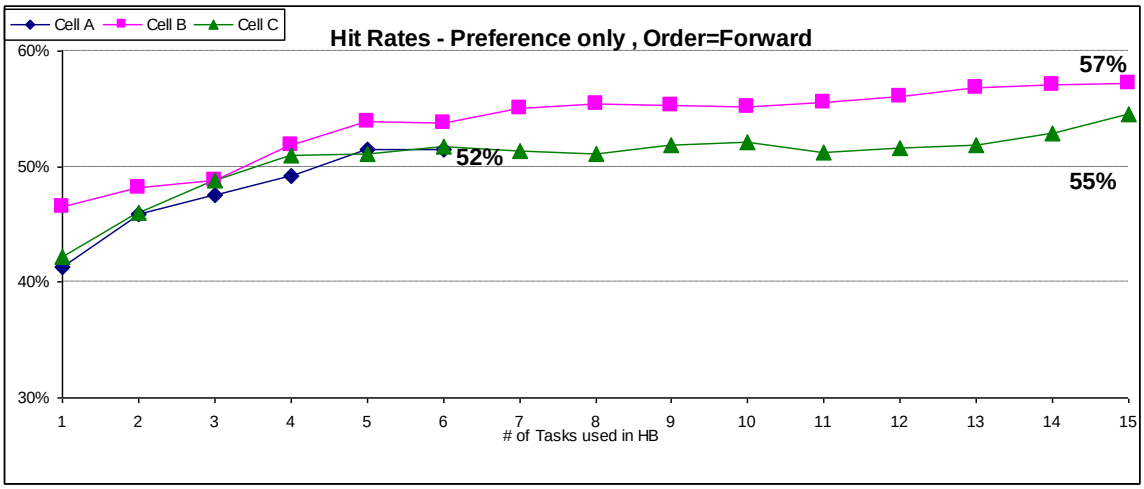
HB models are developed independently for each of cells A, B and C. In the "forward" process, we use all the data up to the i^{th} task, with i=1,2,3,…,6 for cell A, and i=1,2,3…, 15 for

cells B & C. In the "backward" process, all the data based on the last i^{th} task are used to develop the model. In total, 72 HB runs are made. This allows us to compare results from the model based on the first x number of tasks to the model based on the last x number of tasks.

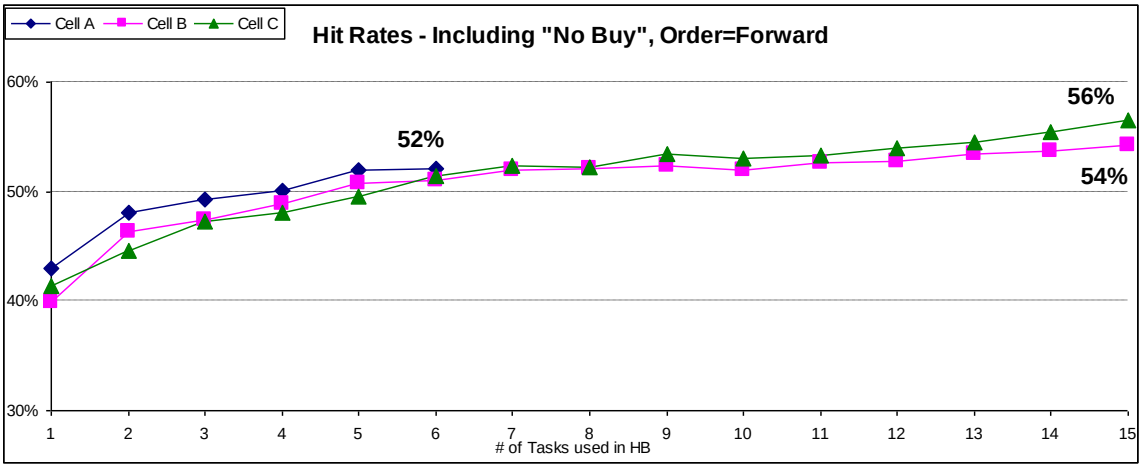
All the models use the first 20,000 iterations as the burn-in period, and the last 10,000 iterations to compute model estimates. No covariates are used in the process and no restrictions are placed on the parameters. Partworth parameters are estimated for all factors and levels as well as a "none" parameter.

In terms of model fit, when all the data are used in each cell, cell A and cell B have comparable RLH statistics (592 vs. 565). RLH is lower for cell C at 463. This is to be expected as respondents are choosing from 5 alternatives in cell C.

We observe that the more tasks used in the HB process, the better we are able to predict the holdout preference from the respondents. The improvements are much larger during the early stages, and tend to level off after about 8 tasks in both cells B and C. Interestingly, cell B has an advantage over cell A right from the start using only the first choice task in the model, and holds that advantage through the next 5 tasks. Cell C, even with more alternatives per task (i.e. more information supplied by each task) performs no better than cell A in the 6 tasks, and lower throughout compared to cell B.



The picture for hit rates when the "No buy" decision is included is less clear. The gaps between the cells are much narrower. Cell A now has an advantage over cell B in the early stages. When all tasks are included, cell C has a slight advantage.



"No buy" is the bottom 3 box of the voting intent question in the dual-response task.

When we compare the simulated respondents' choice to external validity measures, we observe results that are similar to the holdout task. While cell B generally outperforms cell A, the advantage is marginal. While the 15 tasks results in more than doubling of the task length (588 seconds vs. 288 seconds), we have a 7-10% improvement in hit rates. Cell C generally performs the worst by all criteria.

	Approval President[1]	Approval Congress[1]	Vote 2010[2]	Vote 2008[2]	Party affiliation[3]	Political Leaning[3]
Cell A	74%	73%	69%	69%	70%	61%
Cell B	78%	78%	75%	76%	77%	60%
Cell C	72%	72%	68%	65%	75%	59%

1: Hit rate is calculated as an approval rating being equivalent to a choice for a Democratic platform, and a disapproval rating being equivalent to a choice for a Republican platform.

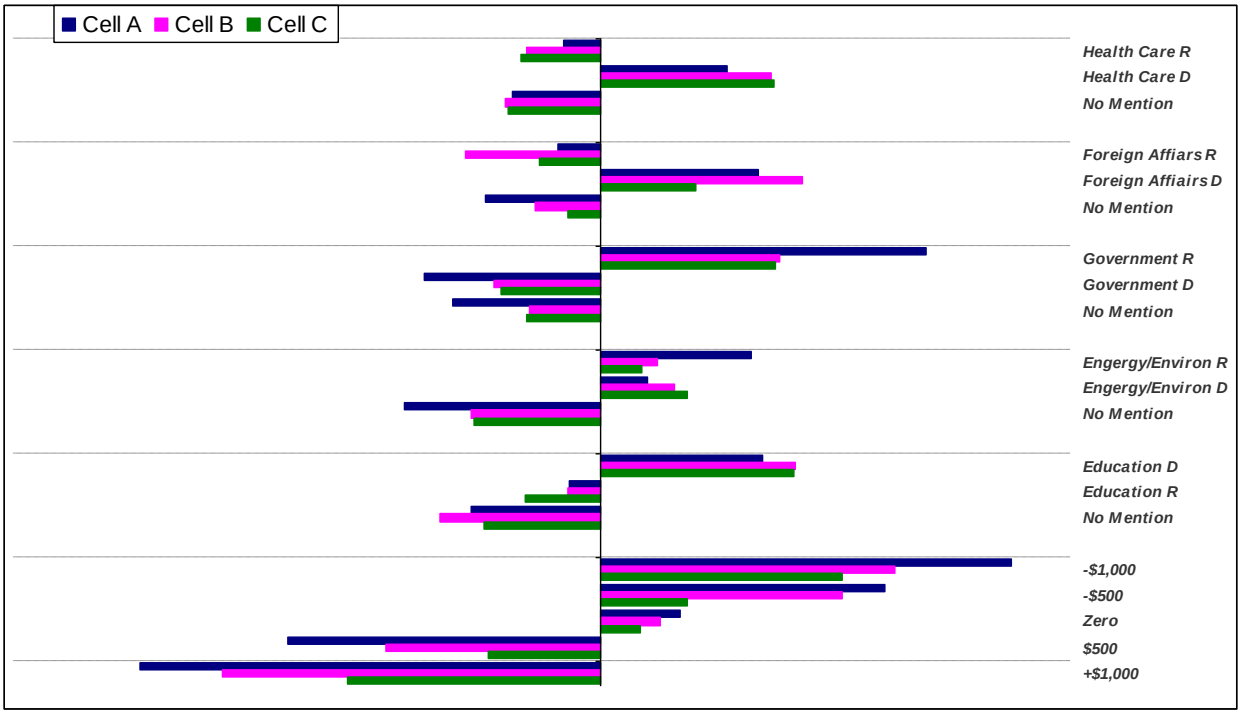
2: Among those who voted for either of the two main parties

3: Exclude those who said "Neither".

In order to understand the impact of party stance in each area, we simulate how likely the respondents are to vote for a candidate with that stance, holding all other areas neutral.

	Option 1	Option 2
Health Care	**Health Care R**	**"No Buy"**
Foreign Affairs	**Neutral**	
Size Of Government	**Neutral**	
Energy/Environment	**Neutral**	
Education	**Neutral**	
Federal Taxes	**Neutral**	

We then center the shares for that candidate across all the levels in each factor to highlight the impact of the various levels. The analysis shows that all the cells generally show similar results with agreement on the directional impact of all factor levels. However, the magnitude of the impacts is much larger in cell A compared to B, suggesting that cell A shows more sensitivity. Cell C shows the least amount of sensitivity, suggesting more noise and lack of consistency in the model.



Another way of assessing model consistency is to examine price reversals. We use the federal tax factor as our pseudo price factor. Logically, all else being equal, paying less tax should be preferred over paying more tax. Since we estimate the five levels in the tax factor as partworth, we can count how often there is a difference in the ordering of the utility levels of the partworth to the logical order.

We illustrate this process with the following examples. While respondent 1260 has the correct order for all five of his partworth, respondent 1799 makes two mistakes in the ordering. Respondent 30896 shows no logical consistency in his partworth, and commits four mistakes.

id	Tax Factor	Beta
1260	Reduce tax $1,000	1.59847
1260	Reduce tax $500	1.12119
1260	No change	0.86754
1260	Increase tax $500	-1.53949
1260	Increase tax $1,000	-2.04771

id	Tax Factor	Beta
1799	Reduce tax $500	1.26765
1799	Reduce tax $1,000	1.20973
1799	No change	0.32314
1799	Increase tax $500	-0.70015
1799	Increase tax $1,000	-2.10036

id	Tax Factor	Beta
30896	Increase tax $1,000	0.80648
30896	Increase tax $500	-0.09276
30896	Reduce tax $500	-0.12834
30896	Reduce tax $1,000	-0.16287
30896	No change	-0.4225

Counting up the mistakes each respondent makes, we conclude that cell A performs best, echoing what we observed in the sensitivity analysis. The additional tasks in cell B result in slightly more mistakes being made. The additional complexity in cell C clearly makes things much worse.

Cell	# of "mistakes"	% all correct	% all wrong
A	1.52	24%	9%
B	**1.69**	23%	12%
C	**2.08**	**12%**	**16%**

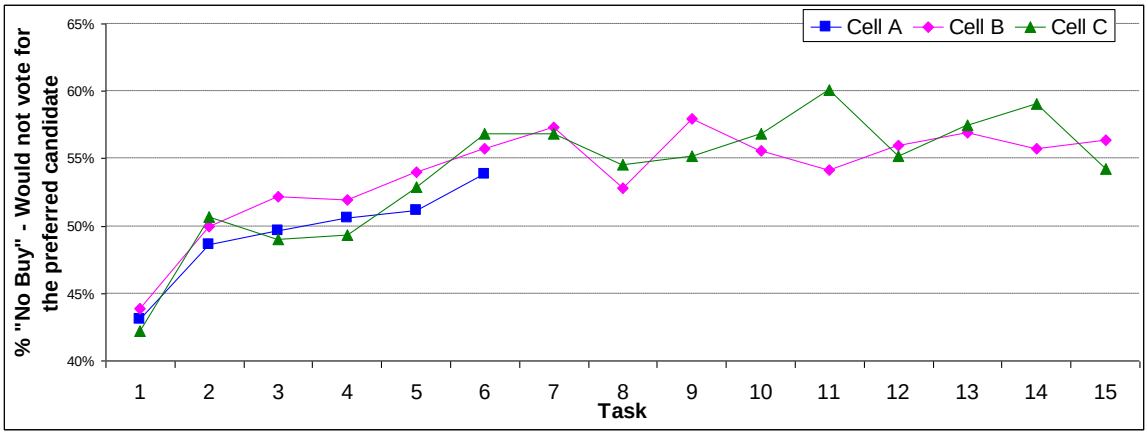
Red/bold indicates significant difference to cell A, **plum/bold** indicates significant difference to both cells A & B, at 5%, two-tailed, ANOVA, Multiple Comparison Bonferroni.

Since cells B and C use more tasks at each respondent level, there is less Bayesian shrinkage towards the overall mean for each respondent. It is possible that this loss of consistency is simply a result of focusing more on the individual and less on the group average (where logical consistency is always better). We repeat this analysis using the HB model developed using only the last 6 tasks in cells B & C to even out the amount of shrinkage towards the mean. The data here shows an even worse deterioration in consistency, suggesting the later tasks are not as "good" as the earlier ones.

Cell	# of "mistakes"	% all correct	% all wrong
A	1.52	24%	9%
B (last 6 tasks only)	**1.96**	**14%**	**16%**
C (last 6 tasks only)	**2.04**	**12%**	**17%**
Red/bold indicates significant difference to cell A, at 5%, two-tailed, ANOVA, Multiple Comparison Bonferroni. No difference between cells B and C.			

Johnson & Orme (1996) saw an increase in selection of "none" during later tasks, but concluded that this was due to the "economic hypothesis" – respondents chose none to indicate that no offering was sufficiently attractive. Suresh & Conklin (2010) linked the increased use of "none" with surveys scored low on respondent engagement.

We again observe this increase of "no buy" decisions in the later tasks. If this is due to the economic hypothesis, we believe increasing the number of alternatives from three to five in cell C should result in a decrease in the amount of "no buy" decisions. We see no evidence of that. Although we have no direct measure of respondent engagement after each task, we believe the increase in the use of "no buy" decisions is likely related to a decrease in respondent engagement in the later tasks.



c. Simplifying Rules

Respondents complete the later tasks much faster. Although this may be largely due to familiarity with the task, Hauser, Gaskin, and Ding (2009) suggest that heuristic simplifying rules may also be at work. We hypothesize that this can take on the form of increased focus on the top factors (most important factors to that respondent) in the later tasks, resulting in less tradeoff and less utility maximization behavior. If compensatory rules are used, a respondent is motivated to maximize utility in his choice, and should choose options closest to his "ideal", with more factors matching his ideal on average across many profiles. When simplifying rules are used, we expect less matching to his "ideal" across multiple choice tasks.

We analyze this by looking at the differences in the models developed using the first six and the last six choice tasks for both cells B and C. Hauser, Gaskin, and Ding (2009) suggest that those who are familiar with the category are more likely to use these heuristics. We compare those who are more familiar with politics against those who are less so to see if we can observe any differences.

First, we look at the individual level factor importance for the top factors. Respondents in cell B placed slightly more importance on the top factors in the later six tasks compared to the first six. The difference is small, but it is there. More importantly, as Hauser, Gaskin, and Ding (2009) suggested, we see a more pronounced difference among those who are considered to be familiar with politics. However, we do not observe this behavior among the cell C respondents.

		Most important			Top 2 Factors			Top 3 Factors		
Cell	Order	Total Sample	Un-familiar	Familiar	Total Sample	Un familiar	Familiar	Total Sample	Un-familiar	Familiar
B	First 6	34.3%	34.9%	33.3%	57.7%	58.4%	56.8%	75.7%	76.2%	74.9%
	Last 6	35.2%	35.5%	34.8%	59.0%	59.2%	58.8%	76.8%	77.1%	76.4%
	Diff.	**-1.0%**	-0.6%	**-1.5%**	**-1.3%**	-0.8%	**-2.0%**	**-1.2%**	-0.9%	**-1.5%**

Red/bold indicates significantly different from 0, at 5%, two-tailed, paired t-test.
No significant difference in cell C.

Before the choice task section, we ask respondents to describe the position of their "ideal" political candidate. This serves to introduce the respondents to the political issues, and also allows us to observe how often the preferred candidate in their choice tasks matches that "ideal". If they identify either party's position as closest to their "ideal" candidate, we count how often their chosen candidate in each choice task has a platform matching their "ideal" candidate's position. If respondents are fully utility maximizing, we expect that on average respondents would choose options that match up as much as possible to their "ideal". In a simplifying decision, we expect less matching between respondent's choice and their "ideal".

We illustrate this matching process with the following example. This respondent's choice in this task matched two out of the four positions (50%) he identified as being closest to his "ideal".

	"Ideal" Candidate	Chosen Candidate
Health Care	The government should provide health insurance for all Americans.	The government should get out of health care
Foreign Affairs	Overseas America should focus on leading the world and promoting our values, and not listen to the UN	Overseas America should focus on leading the world and promoting our values, and not listen to the UN
Size Of Government	The federal government is bloated, corrupt and wasteful—spending needs to be cut dramatically	The federal government is bloated, corrupt and wasteful—spending needs to be cut dramatically
Energy/ Environment		
Education	The best way to improve the education system is by giving more resources to our public school teachers.	

In this analysis, we count the number and percentage matches for the first six choice tasks, and for the last six choice tasks, and calculate the difference.

Again, we see small but significant differences in cell B, showing that respondents do less "ideal" matching in later choice tasks. We observe this difference among those who are more familiar with politics as well as those who are less familiar with politics. We also observe the same behavior among cell C respondents, but curiously, only among those respondents who are less familiar with politics.

		Avg. # of matches			Avg. % of matches		
Cell	**Order**	**Total Sample**	**Unfamiliar**	**Familiar**	**Total Sample**	**Unfamiliar**	**Familiar**
B	**First 6**	**1.81**	**1.74**	**1.92**	**43.9%**	**43.1%**	**45.1%**
	Last 6	**1.74**	**1.67**	**1.84**	**42.1%**	**41.4%**	**43.1%**
	Diff.	**0.08**	**0.07**	**0.08**	**1.8%**	**1.7%**	**2.0%**
C	**First 6**	**1.82**	**1.74**	**1.94**	**47.1%**	**46.4%**	**47.9%**
	Last 6	**1.75**	**1.63**	**1.92**	**45.2%**	**43.0%**	**48.0%**
	Diff.	**0.07**	**0.11**	**0.02**	**1.9%**	**3.4%**	**-0.1%**
Red/bold indicate significantly different from 0, at 5%, two-tailed, paired t-test.							

One might argue that this behavior of less "ideal" matching may be the result of a learned behavior. As a respondent progresses through his choice tasks, his "ideal" may change. This is certainly a possibility. However, if that is the case, we should expect that the choices made in later tasks become more acceptable, resulting in a decrease in the "no buy" decision. As this is not the case, we conclude that respondents are simply less engaged in the process; their later choices are not as good as the earlier ones as shown by the increase in "none" decisions.

Not surprisingly, although respondents in all three cells consider this to be an easy to complete and enjoyable survey, cell A respondents give significant higher ratings in all four feedback questions.

Topbox (5) score, 5-pt scale: 1 - Totally disagree, 5 - Totally Agree	Cell A	Cell B	Cell C
Overall, this survey was easy to complete	66%	**59%**	62%
I enjoyed filling out this survey	59%	**51%**	53%
I would fill out a survey like this again	69%	**61%**	**61%**
The time it took to complete the survey was reasonable	68%	**55%**	**52%**

Red/bold indicates significantly less than cell A, at 5%, one-tailed, one-way ANOVA, multiple comparison (Bonferroni). No difference between cells B and C.

4. Conclusions

Our findings clearly demonstrate the key problem with long choice tasks: respondents become less engaged in the later tasks. Increasing the number of choice tasks brings limited improvement in the model's ability to predict respondents' behavior, and it comes at a cost of deterioration in model sensitivity and consistency.

Respondents behave differently during later tasks. Simplifying behaviors are more likely among those who are familiar with the category. Increasing the complexity of the choice task does not improve any aspect of the model.

We believe our study is the first data point among CBC practitioners showing deterioration in the data quality during later tasks. As this is a very important issue among all market researchers who use CBC data, we intend to replicate this study in other areas.

We recommend that CBC practitioners who utilize on-line panel samples should take pains to keep their panelists and 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, consider increasing sample size (lower sampling error) to compensate for the lower precision in modeling resulting from the smaller number of tasks.

We recognize that market segmentation studies, or any studies that require very precise individual level estimates, may require more tasks. Even in those cases, we must consider the loss of respondent's engagement in later tasks. On one hand, the long choice task section allows us a more precise individual level estimate in terms of predictability. On the other hand, the deterioration in data quality during the later tasks may lead one to question the quality of the estimate in terms of sensitivity and consistency. Future research on this topic is needed to address the best balance between model precision and quality.

REFERENCES

Brazell, J., Diener, C., Karniouchina, E., Moore, W., Severin, V. and Uldry, P. (2006) "The no-choice option and dual response choice designs," Marketing Letters Volume 17, Number 4, 255-268

Hauser, J., Gaskin, S., and Ding, M. (2009), "A Critical Review of Non-Compensatory and Compensatory Models of Consideration Set Decisions," Sawtooth Software Conference Proceedings.

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.

Suresh, N. and Conklin, M. (2010) "Quantifying the Impact of Survey Design Parameters on Respondent Engagement and Data Quality," CASRO Panel Conference.

Tang, J. and Grenville, A. (2009) "Influencing Feature Price Tradeoff Decisions in CBC Experiments," Sawtooth Software Conference Proceedings.

THE STRATEGIC IMPORTANCE OF ACCURACY IN CONJOINT DESIGN

MATTHEW SELOVE
USC MARSHALL
JOHN R. HAUSER
MIT SLOAN

1. STRATEGIC IMPLICATIONS OF ACCURACY IN CONJOINT ANALYSIS

Market simulators based on conjoint analysis help managers predict the market share and profitability of new product designs. Many simulators assume competitors will not respond to a new product introduction. This assumption can lead managers to make poor decisions. For example, Belloni et al. (2008) show that a simulator that ignores competitive response might indicate that the "optimal" new product is identical to a competing product but priced slightly lower. In a real market this design strategy would provoke a price war that would erode the profits of all firms in the market.

To address this problem, academic researchers have created market simulators that account for price reactions by competitors (Choi, Desarbo, and Harker 1990; Choi and DeSarbo 1994; Luo, Kannan, and Ratchford 2007; Luo 2009). These papers assume that competing firms first set non-price features that depend on manufacturing set-ups that can be changed slowly or at high cost. Firms adjust prices relatively quickly and at low cost until all prices reach a Nash equilibrium conditioned on the non-price features chosen in the first stage. Such competitive market simulators provide insight into how a firm's product design choices affect price competition in a market and, hence, lead to more profitable strategic decisions.

In this note we summarize results from a recent working paper where we explore how the accuracy of the conjoint-analysis partworths used in market simulators affects firms' strategic product-design choices (Selove and Hauser 2010). In that paper we address when firms should differentiate their products to soften price competition, and when they offer undifferentiated products. Our key insight is quite relevant to the use of choice-based conjoint analysis: firms acting rationally will make different decisions on differentiation depending upon the amount of noise (or randomness) in customer behavior. This noise is inversely proportional to the logit "scale factor" (Swait and Louviere 1993). We show that when the amount of noise in behavior is small, firms will differentiate their products, but when the amount of noise is sufficiently large, firms will forego differentiation and instead compete for the same customers.

This result has important implications for market researchers. In particular, it implies that firms must accurately estimate the amount of noise in behavior (in other words, accurately estimate the scale factor) in order to develop an optimal product design strategy. If they misjudge the scale factor they will make incorrect strategic decisions and forego potential profit. Many factors contribute to "noise" in real-world choice behavior, including attributes omitted from the conjoint study, changes in behavior across contexts, and inattentiveness or carelessness in customer responses to the survey. If a poorly designed market research study causes respondents to behave more carelessly or randomly than they would in the real world, this will cause a firm to overestimate the amount of noise in choice behavior, causing them to create a product that is too

close to competing products, and thus lead to destructive price competition. On the other hand, if a firm fails to account for sources of noise that exist in the real world, this could lead them to focus too much on differentiating their products even though this means providing less utility to customers.

We illustrate the practical relevancy of these results with an application to a conjoint study on student apparel. We first estimate partworths using CBC/HB analysis, and then hold these partworths fixed while adjusting the scale factor to account for differences in behavior between the initial conjoint task and a hold-out task. (This is similar to the approach suggested by Salisbury and Feinberg 2010.) In the illustrative conjoint-analysis study, accounting for noise across settings implies that a firm should choose the most popular color for its product, even if a competitor also chooses that color. On the other hand, failure to account for this additional noise leads a firm to differentiate its product by choosing a less popular color. This incorrect decision reduces profits for the firm.

2. Intuition for Why Uncertainty Affects Strategic Decisions

Selove and Hauser (2010) provide detailed proofs to illustrate how noise affects product design choices. Although we do not repeat that formal proof here, the intuition can be seen from the following simple example. Suppose a product is available in two colors: grey or red. Most consumers prefer grey, but there is also a segment of consumers who prefer red. Two firms compete in this market, and each firm sells a single type of product. The strategic question is whether, in equilibrium, both firms will choose to produce a grey product (the more popular color), or whether one firm will produce a grey product while the other firm produces a red product to soften price competition. For this example, color and price are the only product features.

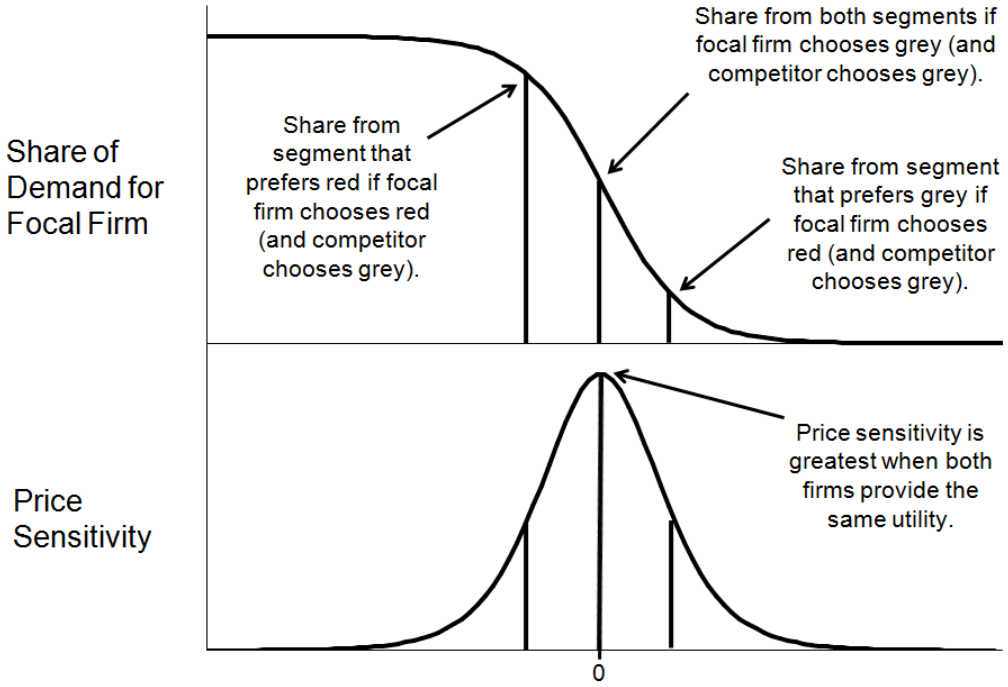
Assume that consumers are described by random utility models where the observed utility component is the standard partworth model (for color and price) and the random component follows a double-exponential extreme-value distribution. In other words, demand follows a logit function. The presence of the random component to utility implies, for example, that some consumers who prefer the color the competitor's product may still choose the focal firm's product (even if prices for both products are the same).

Figure 1 shows how color choice affects competition in the market when there is relatively low utility randomness. This figure assumes (for simplicity of exposition and intuition) that both firms have set the same price and that the competitor has chosen a grey product. The proofs do not require these latter assumptions.

The top half of this figure shows the focal firm's share of demand as a function of the difference between utility provided by the competing firm's product and utility provided by the focal firm's product. The inner bar is for both segments and assumes the focal firm chooses grey; the outer bars are for the two different segments and assume the focal firm chooses red. The bottom half of the figure shows the sensitivity of each segment's demand to changes in price, where "price sensitivity" is the negative of the first derivative of demand with respect to price. Note that the horizontal axis is (for each given segment) utility provided by the competing firm's

product minus utility provided by the focal firm's product.[1] The two firms provide equal utility at the "bar" in the center of the figure.

Figure 1. How Differentiation Affects Demand and Price Sensitivity (For *Low* Randomness in Product Utility)



If the focal firm chooses grey, then each firm receives half of the demand from each segment. Because the derivative of the logit function is highest when the focal firm has one-half of demand, price sensitivity is at its highest at this point. On the other hand, if the focal firm chooses a red product, its share of demand increases for the segment that prefers red, but decreases for the segment that prefers grey. Both customer segments are now less price-sensitive (as is indicated by the two outer bars on the bottom half of the figure, showing lower price sensitivity for both segments). Intuitively, customers who have strong preferences for one color or another are less likely to be swayed by small price differences. Differentiation softens price competition. It is not hard to show that, in equilibrium, prices in the market will be higher. Figure 1 illustrates the basic trade-off faced by a firm choosing red or grey: higher average prices result from each firm choosing a different color, but for the firm that chooses the less-popular color, that color reduces share of demand in the larger segment (while increasing share of demand from the smaller segment). The size of the smaller segment determines whether the net of this tradeoff is profitable for the firm choosing the less-popular color.[2] (We ignore for the purposes of this

[1] To be precise, this axis reflects the difference in *observed* utility, based on color and price.

[2] Although the proofs in Selove and Hauser (2010) assume a logit demand model, the intuition behind these results would hold for any demand function with the property that customers who strongly prefer one firm or another are less price-sensitive. This seems like a reasonable property for demand functions, and it is also consistent with the standard strategic advice that differentiating from competitors helps avoid price wars.

example, which firm gets to be grey and which red. We assume that the focal firm, if it differentiates, is the red firm.)

Figure 2. How Differentiation Affects Demand and Price Sensitivity (For *High* Randomness in Product Utility)

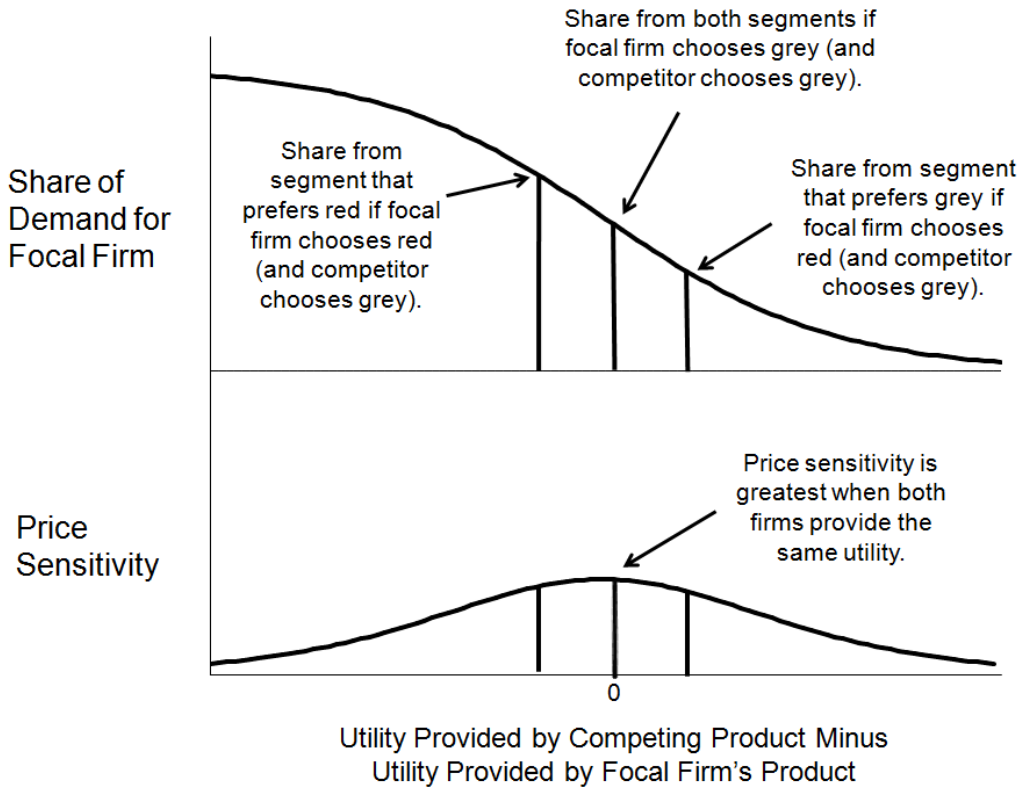


Figure 2 presents the same analysis with one key change: we increase the error term on product utility (that is, decrease the logit scale factor). The demand curve is now flatter, and the first derivative of demand is now smaller. Increasing the amount of noise in customer behavior implies that customers are less sensitive to all product features (including price). In Figure 2 if the focal firm chooses grey, then each firm still receives half of the demand from each segment. As before, if the focal firm chooses a red product, this increases its demand in the segment that prefers red and decreases its demand in the segment that prefers grey. It also reduces price sensitivity for both segments. However, these effects are now smaller. As randomness in behavior becomes greater, a company's color choice has a smaller effect on both demand in each segment and the sensitivity of demand to price.

The arguments so far suggest that the net effect of greater randomness is ambiguous. The reduction in demand due to differentiation (choosing a red product) is less when randomness increases, but so is the softening of price competition. However, the additional randomness makes customers less sensitive to changes in price, hence equilibrium prices (and profit margins) are higher, all else equal. This implies that even a small decrease in demand has a substantial effect on profits, since this small change is multiplied by larger profit margins. Therefore, although the net benefit from differentiation (softening price competition) becomes trivial as

randomness increases, the cost of choosing a less popular color is still substantial. When randomness becomes sufficiently high, this confluence of effects leads all firms to choose the most popular color (grey) in equilibrium.

To summarize, two firms face the standard differentiate-or-not dilemma – greater differentiation reduces price competition but less differentiation allows both firms to focus on the highest-demand segments. The key new insight is that inherent uncertainty in consumers' choice behaviors affects how firms resolve the dilemma. Greater uncertainty leads to less differentiation.

This insight has roots in prior research that suggests similar results for a demand model in which preferences are uniformly distributed, all products have the same marginal cost, and all customers have the same price sensitivity (de Palma et al. 1985; Irmen and Thisse 1998). One contribution of our paper is that it extends this result so that it is connected to the means by which firms measure consumer preferences—choice-based conjoint analysis. This connection is critically important because it allows us to demonstrate that the accuracy of market research, not just the relative partworths or the distribution of relative partworths, will determine how firms make strategic decisions on differentiation.

Specifically, the "scale factor" in choice-based conjoint analysis (the logit model) is a function of inherent uncertainty and uncertainty due to measurement. By inherent uncertainty we mean residual uncertainty (stochasticity) in consumer decisions, stochasticity that might be due to actions beyond the firm's control. Uncertainty in measurement is a function of the quality of the questionnaire, the completeness of the set of product features, and the ability of an estimation method to uncover accurate parameters from the data.

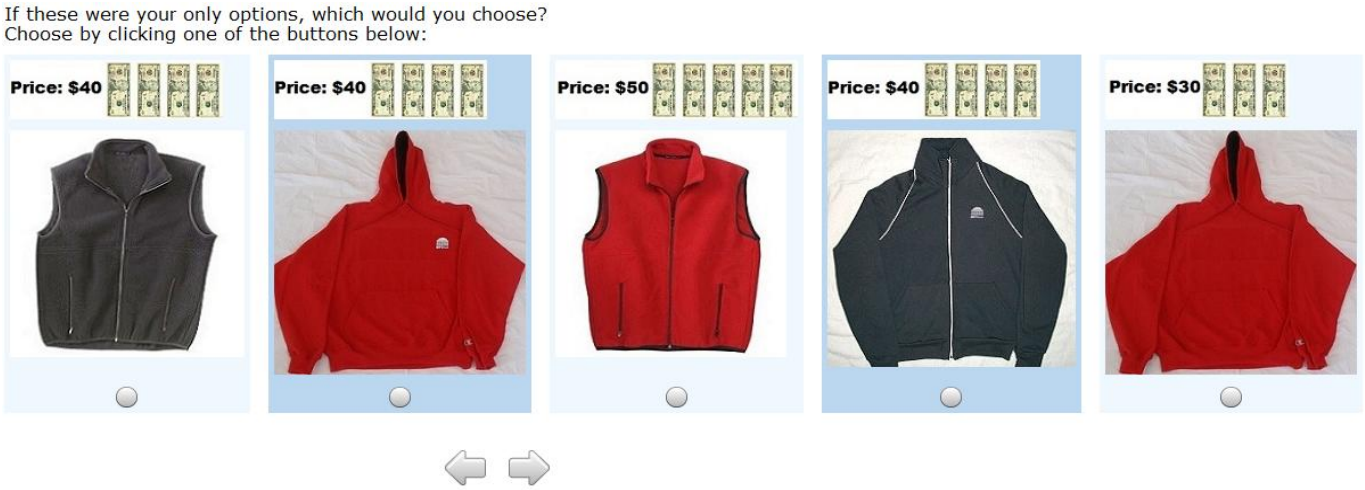
3. Illustrative Example: Conjoint Study on Student Apparel

Selove and Hauser (2010) illustrate the practical impact of the theoretical result with an example drawn from a conjoint study on student apparel. Thirty-eight students completed a CBC study with the following features. The study had four cells that varied in a 2 x 2 design of {careful design and graphics vs. less careful design, words only} x {incentive compatible vs. not incentive compatible}. For the purposes of this illustration we focus on the nineteen students in the careful-design-and-graphics cells. The product features and levels were:

- Type of clothing: track jacket, sweatshirt, or fleece vest
- Color: grey or red
- Logo: School logo or no logo
- Price: Base level ($30 for the sweatshirt; $40 for the other two); or Base level plus $10

Figures 3 presents a sample choice set. To keep the illustration simple, we have foregone a "no choice" option.

Figure 3. Sample Conjoint Question



Each respondent answered 16 conjoint questions, then, after several memory-cleansing tasks, completed a hold-out task in which they ranked their top 5 out of 12 products. (Asking respondents to rank five products instead of choosing one increases the statistical power of our validity tasks.)

Table 1 reports average partworths computed using Sawtooth's CBC/HB software. The data are from the initial conjoint task. On average, respondents prefer the track jacket, the color grey, the school logo, and lower prices.

Table 1. Average part-worth for each feature level

Sweatshirt	0.0
Fleece vest	-0.7
Track jacket	4.9
Red	0.0
Grey	1.5
No logo	0.0
School logo	4.5
Base price	0.0
Plus $10	-2.7

Following Salisbury and Feinberg (2010), we adjust the scale factor to account for variance in behavior between the original conjoint task and the hold-out task. This enables us to parse the random component of consumers' utility functions into components: (1) randomness in the ability of the conjoint model to predict behavior in the calibration choice setting and (2) randomness that accounts for changes in behavior across settings. The scale factors estimated by the original HB partworths account for the first type of randomness. Comparison of predictions to the validation task account for the second type of randomness.

To estimate the second component of randomness, we hold fixed the partworths estimated from the calibration data and then estimate an adjustment to the scale factor using the hold-out data. Our estimates suggest that that the scale factor needs to be adjusted downward by 0.65 to account for the second type of uncertainty.

We next compute equilibria in a simple product-design game. Two firms each produce a track jacket with a school logo. Firms simultaneously choose colors, then simultaneously set prices. Firms make their color decisions based on (possibly inaccurate) market research, and cannot later change the color of their product once they observe demand. However, prices can be easily adjusted and will reach a Nash equilibrium based on the "true" model of customer behavior. In this game scenario, firms that conduct inaccurate research might make sub-optimal color choices due to inaccurate predictions of demand and of equilibrium prices. In light of our earlier theoretical arguments, firms with inaccurate market research might make erroneous decisions to differentiate and then face a "world" in which differentiation might not have been the better strategic solution, or vice versa.

To demonstrate why market research has strategic implications, we assume that the "true" model of customer behavior is represented by the partworths as estimated on the calibration data, but adjusted by a factor 0.65 to represent the second form of uncertainty. If both firms know the true model of behavior, the best equilibrium strategy is for both to produce a grey jacket. Equilibrium mark-ups are $11.40 per jacket, and each firm captures one-half of the potential customers. They each have equilibrium profits of $108.30 among the nineteen students.

Now assume one firm believes that its market research is more accurate than it really is. We represent these delusional beliefs by assuming the firm fails to adjust the scale factor to account for the uncertainty between calibration and validation. Under these conditions the delusional firm (incorrectly) predicts that differentiating its product will increase its equilibrium earnings. It does so but is surprised when true demand is at variance with its predictions. Its actual earnings are lower than they would have been by 3.4%. (The magnitude is not important; this example is illustrative. Different assumptions on factor costs could make this percentage much larger.)

Table 2 provides more detail on equilibrium prices and profits assuming that market research accounts for both forms of uncertainty accurately. Although Firm B would prefer a differentiated market, this is not an equilibrium. Firm A is faced with the differentiation dilemma and should resolve it toward no differentiation (because uncertainty is high). The only equilibrium is the undifferentiated market where both firms offer grey jackets.

Table 2. Equilibrium Prices and Profits in the "True" Market

	Firm A	Firm B			Firm A	Firm B
Color	Grey	Grey		Color	Red	Grey
Marginal Cost	$40.00	$40.00		Marginal Cost	$40.00	$40.00
Equilibrium Price	$51.40	$51.40		Equilibrium Price	$52.63	$56.31
Profit Margins	$11.40	$11.40		Profit Margins	$12.63	$16.31
Demand	9.5	9.5		Demand	8.3	10.7
Profits	$108.30	$108.30		Profits	$104.64	$174.73

Table 3 provides detail on a simulator that Firm A would use if it did not recognize the need to adjust the scale factor. Firm A underestimates the true uncertainty in the market and resolves the differentiation dilemma in favor of differentiation. It predicts that both firms are better off differentiating and, thus, produces a red jacket.

Table 3. Firm A's Simulator Based on Inaccurate Market Research

	Firm A	Firm B			Firm A	Firm B
Color	Grey	Grey		Color	Red	Grey
Marginal Cost	$40.00	$40.00		Marginal Cost	$40.00	$40.00
Equilibrium Price	$47.35	$47.35		Equilibrium Price	$49.54	$53.18
Profit Margins	$7.35	$7.35		Profit Margins	$9.54	$13.18
Demand	9.5	9.5		Demand	8.0	11.0
Profits	$69.86	$69.80		Profits	$76.08	$145.32

In this example, Firm A is pleasantly surprised when it launches its product because it actually earns $104.64 rather than $76.08. The market research firm is probably rewarded and rehired. However, Firm A does not observe the opportunity loss because its "but-for" world is not accurate. Firm A never knows that it could have earned even greater profits by launching a grey jacket. (However, there are some hints. If Firm A knew that uncertainty implied lack of differentiation, it should be suspicious because its simulator under-forecast profits by 27%.)

We find this example compelling. Firm A makes the wrong strategic decision because it is unaware that its market research is inaccurate. More importantly Firm A never gets to observe its opportunity cost and is pleasantly surprised by the market outcome. Firm A will continue to rely on inaccurate market research and continue to make incorrect strategic decisions. This example illustrates why it is imperative that a firm adopt best practices in market research. The example also motivates academic research for improved measurement and estimation. Even small improvements might tip the scales in the differentiation dilemma.

This example also illustrates the pitfall of relying on internal validation only. It is not hard to image a "quick-and-dirty" CBC study that has good internal validity but poor external validity. For example, the stimuli might make a feature unnecessarily salient, key features may be left out of the mix, or the questions might be worded incorrectly.

Our example is illustrative. Our "validation task" is a within-respondent holdout task and, hence, may not capture all type-2 uncertainty. Nonetheless, by extension of our arguments, firms are advised to undertake true external validity studies. Such studies may pay off by assuring that the firm makes the right strategic decisions.

4. Conclusion

Recent academic research has developed tools that enable managers to predict how product-design decisions affect price competition in a market. Selove and Hauser (2010) show that a firm's strategic behavior can depend upon the accuracy with which it predicts consumer behavior. As true noise in behavior becomes greater, firms shift their emphasis away from product differentiation and focus on the largest segment of demand. Incorrect estimates of the logit scale parameter lead to costly strategic errors.

It is important to conduct market research studies that accurately represent the level of care and thought involved in real-world decision-making, for example, by providing incentives for truthful responses (Ding, Grewal, and Liechty 2005; Ding and Huber 2009), by making sure respondents are familiar with the attributes and the task (Johnson and Orme 1996), and – as was the focus in the current paper – by adjusting for variance in behavior across settings (Salisbury and Feinberg 2010). Such procedures help ensure that managers neither overestimate nor underestimate the true amount of noise in behavior and, hence, make the correct strategic decisions.

REFERENCES

Belloni, A., R. Freund, M. Selove, and D. Simester (2008), "Optimizing Product Line Designs: Efficient Methods and Comparisons", *Management Science*,.54, No.9, p. 1544-1552.

Choi, S.C. and W.S. DeSarbo (1994). "A Conjoint Simulation Model Incorporating Short-Run Price Competition," *Journal of Product Innovation Management*, 11, p. 451-459.

Choi, S. C., W. S. Desarbo, and P. T. Harker (1990). "Product Positioning under Price Competition," *Management Science*, 36, 2, p. 175-199.

de Palma, A., V. Ginsburgh, Y. Y. Papageorgiou, and J. F. Thisse (1985), "The Principle of Minimum Differentiation Holds under Sufficient Heterogeneity," *Econometrica*, 53, 4, p. 767-781.

Ding, M. (2007). "An Incentive-Aligned Mechanism for Conjoint Analysis," *Journal of Marketing Research*, 42, 2, p. 214–223.

Ding, M., R. Grewal, and J. Liechty. (2005) "Incentive-Aligned Conjoint Analysis," *Journal of Marketing Research*, 42, 1, p. 67-82.

Johnson, Richard and Bryan Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" Sawtooth Software Research Paper.

Luo, L. (2009). "Product Line Design for Consumer Durables: An Integrated Marketing and Engineering Approach," Mimeo.

Luo, L., P. K. Kannan, and B. T. Ratchford (2007). "New Product Development Under Channel Acceptance," *Marketing Science*, 26, 2, p. 149–163.

Salisbury, L. and F. Feinberg (2010). "Alleviating the Constant Stochastic Variance Assumption in Decision Research: Theory, Measurement, and Experimental Test," *Marketing Science*, 29, 1, p. 1-17.

Selove, M. and J. R. Hauser (2010). "The Strategic Importance of Accuracy in Conjoint Design," Working Paper.

Swait, J. and J. Louviere (1993). "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models," *Journal of Marketing Research*, 30, 3, p. 305-314.

PRODUCT PORTFOLIO EVALUATION USING CHOICE MODELING AND GENETIC ALGORITHMS

CHRISTOPHER N. CHAPMAN
MICROSOFT
JAMES L. ALFORD
BLINK INTERACTIVE

ABSTRACT

We describe using genetic algorithm (GA) models to find near-optimal product portfolios in the presence of competition, using individual-level part worths from choice-based conjoint (CBC) and adaptive CBC (ACBC) models. We describe how to find optimal product portfolios, inform portfolio size, and generate hypotheses about product opportunities. Optimization routine is probabilistic and subject to data limitations and overfitting, so we bootstrap the process to find the expected distribution of likely outcomes across many resampled runs. We view this as an exploratory process to generate hypotheses in a product space; it is not a confirmatory or probative method. We offer the computer code necessary to run the GA portfolio model in R, given individual-level part worth estimates from another source (such as CBC/HB or ACBC).

INTRODUCTION

Like many marketing research organizations, at Microsoft Hardware we regularly collect information from discrete choice models to inform product design, engineering tradeoffs, and pricing. We routinely use choice-based conjoint analysis (CBC) and adaptive choice-based conjoint (ACBC) to inform our design decisions and category strategy. Given the success of those projects (e.g., Chapman et al, 2009; Chapman, Alford, and Love, 2009), we wondered whether conjoint analysis (CA) information could be used to inform higher-level strategic questions, such as whether we are making not only individually optimized products but also the optimal *number* of products within a category.

The general issue we examined was this: if we had insight into an entire portfolio – namely, the entire group of products that comprise a firm's offering in a category – what could we do with that information? These questions include: Are we making the right products? Are we offering too many products within a category (or not enough)? Are we differentiating our products effectively within the category? Are there potentially appealing products that we are not offering?

We sought a method that would help answer such questions. These issues are ultimately strategic in nature, so we did not seek to offer a conclusive empirical answer. Rather, we sought to develop a data-driven and reusable process that would inform strategy from an empirical point of view, and that could identify areas for further exploration and consideration for strategy and research. Additionally, we desired to design reusable and freely-available code so other analysts could explore the same questions with their data sets.

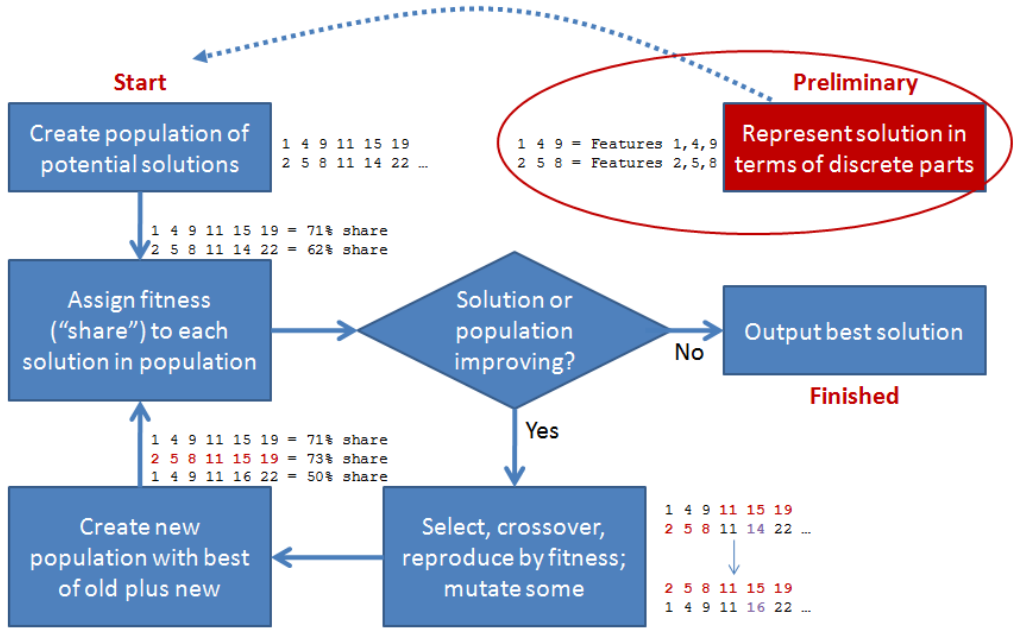
METHOD: OVERVIEW

To evaluate an existing product portfolio, we need a basis for comparison. A natural basis is to contrast the portfolio to an "ideal" portfolio. But how can one find an ideal portfolio? One possibility is to search the space of possible portfolios, and to evaluate each possible portfolio for customer appeal. To do this, one needs a method to search the portfolio space, which is likely to be large and complex, to evaluate each candidate portfolio, and to iterate towards an optimal solution.

Our process involved describing a candidate portfolio as a set of products, where each product comprises a set of attributes and features. Each candidate portfolio ("OURS") is evaluated against a set of other products ("COMP") that consists of current and anticipated competitive products. For each candidate portfolio we determined how many respondents would choose some product from OURS rather than one from COMP (or "None"). Each individual's preference within the portfolio was determined using individual-level HB part worths.

For a product space of any substantial complexity, there are too many portfolio possibilities to investigate through exhaustive search; a heuristic search method is needed to make the process computationally feasible. Belloni et al (2008) demonstrated that a genetic algorithm (GA) model is able to find near-optimal solutions to portfolio preference problems. We implemented such a GA method to select and optimize candidate portfolios. In that GA process, a population of potential portfolios is created initially at random; is evaluated and recombined to yield a new population; members of the population are selected according to fitness, mutated and/or recombined to yield a new population; and this iteration and recombination process is continued until additional improvement is unlikely. Figure 1 shows an outline of the GA process; we discuss the details of each stage below.

Figure 1:
Schematic of GA Process (for a single run among many bootstrapped iterations)



With a single sample of data, GA models may overfit and capitalize on chance within the dataset. To counteract this, we implemented a bootstrapping approach, where a subset of respondent data is used to find a single "near-best" portfolio with one complete evolution of the GA search procedure (and, optionally, preference is itself bootstrapped within each evaluation cycle internal to the GA model). This process is conducted many times for different samples of respondent data, with each proposed near-best portfolio evaluated against holdout respondents not used in that iteration of the GA search. We then examine the distribution of results in the holdout evaluations, where we inspect the total portfolio result compared to other variables such as portfolio size, distribution of price points, features offered, and match to existing product offerings.

A single near-best portfolio may be obtained from the Sawtooth Software Advanced Simulation Module (ASM; Sawtooth Software, 2003). However, ASM does not run repeated samples with varying respondent sets and holdout respondents. Thus, to use ASM to examine the distribution of likely outcomes would require substantial manual work to resample respondents, re-run a model, and compile results. We believe the current version of ASM is especially valuable for single-product optimization and for basic exploration of portfolio models; depth exploration of portfolio distributions may be more easily conducted with a bootstrapped model as is presented here.

Detailed Method

Genetic representation of portfolio

The first stage is to design a "genetic" representation of a possible solution, namely, a representation in which each functional element is represented as a discrete, replaceable part similar to a gene (Goldberg, 1989). In the present case, a single genetic solution is a portfolio of products, where each product is a list of features and attributes. Table 1 defines a product space with three attributes with 3-4 levels each.

Table 1: Product Attributes and Levels for a Simple CBC (example)

Attribute	*Levels*
1	1, 2, 3, 4
2	1, 2, 3, 4
3	1, 2, 3

For instance, "Attribute 1" might represent the feature "size," which occurs in four levels, while Attribute 2 could be "price," and so forth. A complete product is specified by choosing one level for each attribute.

A portfolio, then, is a collection of differentiated products specified in terms of their feature levels. Table 2 shows a possible lineup of products given the attributes and levels shown above.

Table 2: A Hypothetical Portfolio

Product	*Attribute 1*	*Attribute 2*	*Attribute 3*	*Etc.*
Product 1	Level 2	Level 1	Level 2	…
Product 2	Level 1	Level 1	Level 1	…
Etc. …				

After fielding a CA study, we typically have individual-level part worth estimates of the importance of attribute levels for each respondent. Those map directly to the list of attribute levels in successive columns within the part worth data, as shown in Table 3.

Table 3: Map of Features to Part Worth Data Columns

Attribute	*Levels*	*Part worth columns*
1	1, 2, 3, 4	1, 2, 3, 4
2	1, 2, 3, 4	5, 6, 7, 8
3	1, 2, 3	9, 10, 11

Each product is represented by the numbers of the columns that correspond to its feature levels. For instance, "Product 1" from Table 2 comprises features 2, 1, and 2 for its attributes, respectively, and those levels are represented by columns 2, 5, and 10 in the part worth data. Table 4 presents the column representation for the portfolio above.

Table 4: Portfolio Representation as Column Positions

Product	*Part worth columns*
Product 1	2, 5, 10
Product 2	1, 5, 9
…	

For the purpose of the GA model, this may be compacted into a single string in which successive products are simply compiled in order, as shown in Table 5.

Table 5: Genome Representation of Portfolio

Portfolio 1:	2, 5, 10, 1, 5, 9, …

With this representation, a portfolio is specified as a simple vector of integers. In this example, there are 3 integers per product such that a portfolio with 8 products would require 24 integers that represent the corresponding columns in the part worth data set.

Assessment of portfolio fitness

The GA method operates by finding a genetic representation – in this case, a vector of integers – that represents the best fit to some "fitness function." The key requirement for the analyst is to write this function for the problem at hand. Useful functions could be things such as the absolute preference that respondents have for a portfolio (i.e., likelihood to choose a product from it), or variants of preference such as maximized share vs. competition, revenue, or profit.

In the present study, the fitness function was designed to estimate the proportion of respondents who would choose any product from the portfolio, as opposed to choosing none. To estimate this, we evaluate the share of preference for each individual for each product, using the standard multinomial logit (MNL) model. The estimated "none" part worth is included. Each respondent is assigned by strict first choice to "prefer" the product (or "none") that receives the highest summed part worth score for its attributes. The fitness function then returns the proportion of respondents who choose any of the products in the portfolio as opposed to none. (This may be extended easily to consider competitive products by including them in a non-varying and non-evolving portion of the portfolio genotype.)

From the analyst's point of view, the most complex portion of the portfolio GA is writing and testing the appropriate fitness function that correctly evaluates a candidate portfolio and returns its value. Careful attention must be given to correct implementation of the MNL model and to appropriate pricing and feature prohibitions or interactions.

Genetic algorithm parameters

Given an appropriate fitness function, standard GA software may be used to find candidate solutions. We used the "rgenoud" package for the R statistics environment (Mebane and Sekhon, 2009; R Core Development Team, 2010), with parameters as detailed in Table 6.

Table 6: GA Parameters

GA library:	rgenoud (Mebane and Sekhon, 2009)
Genome structure:	1 integer per attribute (min=1, max= # of levels) * # attributes * portfolio size.
Population size:	400
Maximum generations:	50-200 (variable according to fitness trend)
Elitism:	Yes (best candidate always preserved)
Operators:	cloning; simple crossover; heuristic crossover; uniform mutation; boundary mutation; non-uniform mutation; and whole non-uniform mutation. (applied with equal odds across all reproduction events)

It is helpful to experiment with GA parameters in pilot runs. Initial exploration with our data set showed that improvement occurred rapidly within the first 50 generations of the GA model, so we specified 50 as the target number of generations unless recent improvement was shown. Likewise, experimentation with various population sizes showed that we needed at least 200 but no more than 600 population members, so we settled on a standard size of 400 members. Each member represents one complete, proposed portfolio.

Data sets, model iteration and bootstrapping

We used two datasets for this project: a CBC data set with N=716 respondents and an ACBC set with N=405 respondents. The attributes and features varied slightly between the CBC and ACBC data sets but concerned the same product category and had mostly identical feature levels. The CBC data had a total of 8 attributes with 34 feature levels, while the ACBC data had 9 attributes with 29 total levels. In our findings below, we compare the CBC and ACBC results for portfolio size, and then alternate consideration of the CBC and ACBC data for illustrative purposes.

The individual-level part worths came from CA studies implemented using Sawtooth Software SSI/Web (Sawtooth Software, 2010) with hierarchical Bayes (HB) estimation. The CA

surveys were administered online to adult respondents in the US through a third-party panel provider. Individual-level part worths were estimated using Sawtooth Software CBC/HB and ACBC, respectively. For this present analysis, each individual's part worths were assigned to his or her within-respondent mean HB beta estimates. (Using individual HB draws instead of the respondent mean is an option in the available software code.)

For each data set (CBC and ACBC), we investigated different possible sizes of portfolio ranging from k=1 to k=20 products (specifically, k=1,2,4,6,8,10,12,16,20). At each size of portfolio, we run 50 iterations of the GA model to find 50 potential best portfolios. In those iterations, 60% of the respondents' individual-level part worths were sampled and used to evolve the GA, while 40% of the respondents were held out and used to evaluate the final result from the GA iteration.

For each proposed optimum portfolio, we recorded the portfolio size (number of products), the list of products and feature levels, and the preference share for each product and for none. The R statistics environment (R Development Core Team, 2010) was used as the analytic package after importing HB estimates that were saved as a CSV file in Sawtooth Software SMRT. Generic GA functions were provided by the rgenoud package (Mebane and Sekhon, 2009), while the customized fitness function and utility code was written by the first author.

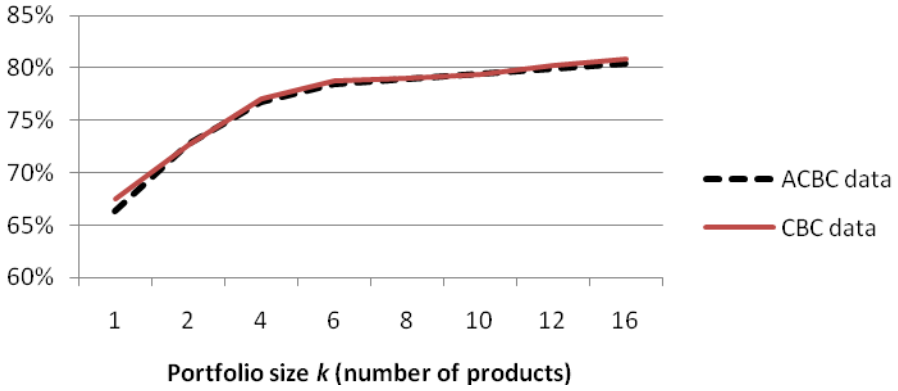
FINDINGS

Portfolio size

Given the product attributes, how many products are needed to satisfy consumer demand? We addressed this by examining the proportion of people who would choose something other than "none," as portfolio size increased from 1 to 20 products (sampled 50 times for each portfolio size).

The median results using CBC and ACBC data appear in Figure 2, while the empirically observed credible intervals for CBC are shown in Figure 3.

Figure 2: Median preference share met, by portfolio size

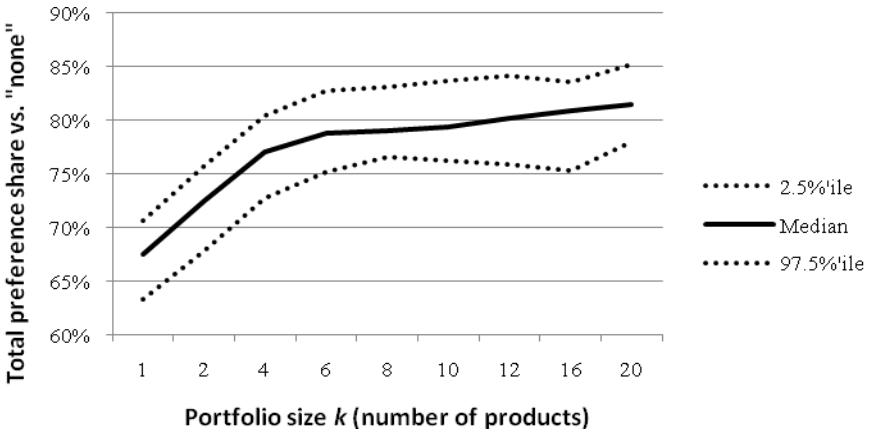


As shown in Figure 2, the incremental gain in preference share levels off sharply after k=6 products. Adding differentiated products continues to increase total preference, but each

additional product above k=6 accounts for less than 1% increment in total-portfolio preference share.

It is striking that CBC and ACBC data show a virtually identical pattern of preference share by portfolio size; this gives a strong validation of the result with regards to internal model consistency.

Figure 3: 95% empirical credible intervals for preference share by portfolio size (CBC data)



In Figure 3, we see the observed credible intervals of preference share with L=50 runs at each portfolio size. The 95% intervals are approximately +/- 3% in preference share at each portfolio size. Above K=6 products, the lower bound of confidence never crosses the median estimate for K=6. This strengthens the observation that additional products are unlikely to satisfy substantial incremental demand (given the attributes and features studied).

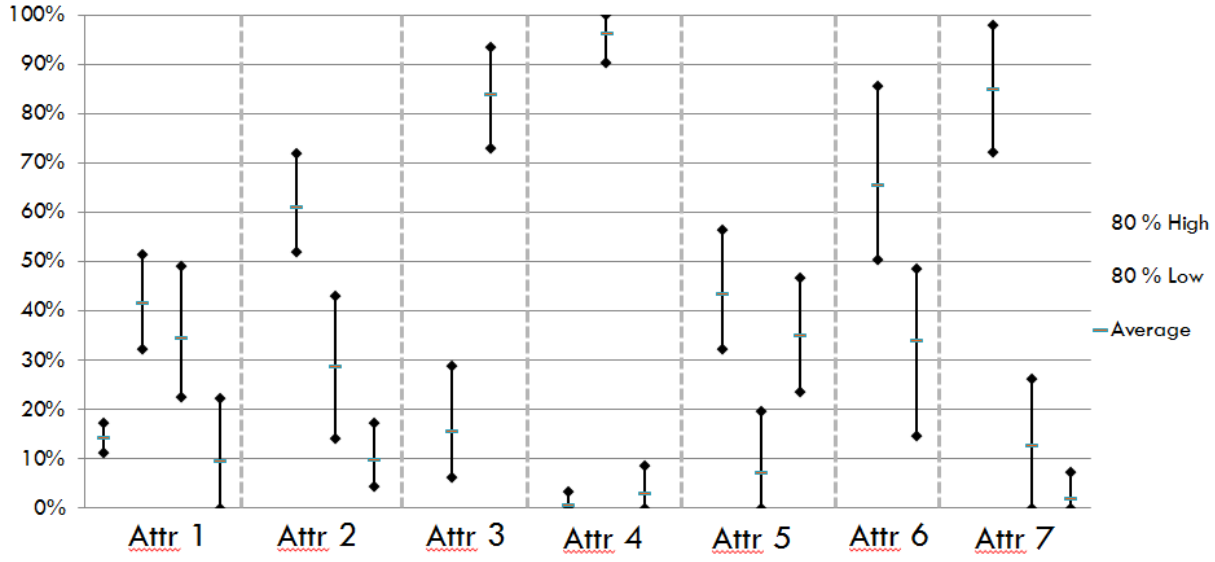
Within-portfolio preference share by feature

The portfolio data may be sliced by feature to examine feature demand. Although feature demand can be calculated quite easily from part worths using the multinomial logit (MNL) model, determining it on the basis of whole-portfolio demand has some advantages. First, feature co-occurrence can be examined easily (driven by respondent preference patterns). Second, the demand is bounded by portfolio size and thus may be forced to 0% or 100% more often than would be observed in part worths alone; this can be salient for managerial purposes. Third, it provides an alternate method of feature estimation in the presence of competition and boundaries, which may validate MNL estimation or provide another point of comparison to market data.

Feature level preference may be computed by examining each portfolio for each feature, and summing the preference of all the products within a portfolio in which that feature exists. For instance, suppose that in a k=6 product portfolio, 3 products have feature X. If the estimated demand of those 3 products is 5%, 4%, and 16%, respectively, then feature X would have an estimated demand of 5+4+16 = 25%.

Figure 4 shows 80% credible intervals for demand by feature level, across L=100 runs of portfolios with k=6 and k=8 products (ACBC data).

Figure 4: Preference share by feature in k=6-8 product portfolios



In Figure 4, we see that many features have quite wide ranges of estimated demand, indicating that they are of interest to consumers but may be managed with relatively great latitude within a portfolio. Other features, however have almost zero demand (e.g., Attribute 4-Feature 1) or universal demand (Attribute 4-Feature 2).

In addition to investigating demand, we have compared these estimates to actual market data. Those findings cannot be reported in detail due to confidentiality and market data restrictions, but one comparison may exemplify them. Attribute 2-Feature 2 is a relatively unique feature whose market demand is of substantial interest to our product team. Its current market penetration is approximately 35% (+/- 4%) according to a recent, separately fielded survey by our team. We see in Figure 4 that Attribute 2-Feature 2 has an estimated portfolio-basis demand of approximately 14%-43%, with a median estimate of 29%. The median estimate is quite close to the actual market penetration and the range overlaps the actual value; thus the portfolio estimate is consistent with performance that we should expect in the market.

This kind of estimate may be useful as a diagnostic indicator. For instance, if a feature were performing worse in the market than was expected in the portfolio model, one could consider targeting it for increased communication or other market intervention.

Product opportunities

To find potential opportunities, one may examine the products that often appear across the optimal portfolios. A simple way to do this is to count the number of times that a fully-specified product (i.e., a complete product string) appears across portfolios.

Table 7 lists the products that appeared in more than 20% of CBC portfolio runs (excluding runs with K=1 or 2 products). Seven products appeared very commonly, of which two are currently not available in the market (lines 3 and 6 in Table 7). Comparing those products to the others, the unique feature of those two products is the combination of attribute 2/level 2 (“x2xxxxx” in the feature code list) occurring with attribute 6/level 2 (“xxxxx2x” in the list).

Table 7: Proportion of times a given product appears in a portfolio with K≥4 products (currently unproduced products shown in bold)

	Proportion of all portfolios (N=800, K≥4)	Feature codes (excluding price)
1	0.76	2111112
2	0.47	1311512
3	**0.45**	**3211422**
4	0.26	1121512
5	0.23	2111111
6	**0.22**	**3211122**
7	0.21	3111412

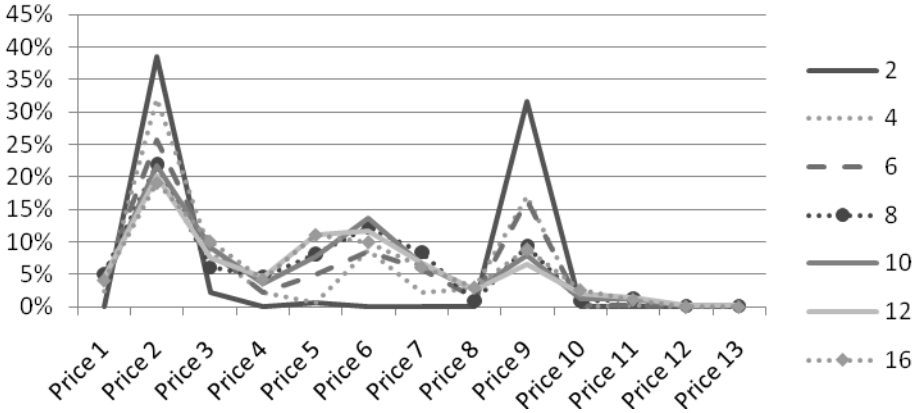
We used this information to investigate the feasibility and cost of combining those attributes in a new product. More generally, we find that this type of investigation can be an easy way to generate product ideas with existing data. Those ideas must be subjected to further vetting and confirmation, but we believe it is useful to have an automated procedure of this kind to generate ideas along with initial supporting data.

Price bands

The fitness algorithm requires pricing information (if relevant) to determine the portfolio composition. By inspecting the price distribution of products found in optimal portfolios, one may form hypotheses about consumer price expectations.

Figure 5 shows the occurrence of products in optimal portfolios, counting the occurrence of products at given price points. Price 1 to Price 13 span the common (but not entire) range of product pricing in the category of interest. When only two products are produced, the most common price points are Price 2 and Price 9. As more products are produced, the distribution of price points becomes more diffuse, yet there are still obvious peaks around Price 2, Price 6, and Price 9.

Figure 5: Product distribution by price and portfolio size (K=2-16, ACBC data)



These results raised several ideas for the product team. First, they call into question a highly stacked pricing approach that attempts to offer product at many different price points. The results suggest that it may be better to offer products at fewer price points but with more differentiation (e.g., along the lines of the opportunities identified in the previous section).

The findings suggest that there is little demand for very high-priced products in this category (prices above Price 9) when the portfolios are otherwise near optimal. Thus, the existence of high-priced products in the market may be due to inefficiencies in portfolio offerings rather a consumer desire for fully-loaded products as such.

The results also suggest that most consumers are willing to pay slightly more than Price 1 if they are offered a feature of interest. This information was of great interest to retail partners. Again, the information is exploratory, not definitive, but is nevertheless useful because it is easily available from this process applied to existing data, while it might be difficult or impossible to obtain otherwise.

DISCUSSION

There are several limitations of this method and areas for exploration. The primary limitation is that this process is only as good as the data provided. It assumes that one has collected information correctly, with an appropriate sample and CA design, for attributes that accurately and completely define a product space. None of those is a simple requirement. If important attributes are omitted, then the results will be difficult to interpret at best, or misleading at worst. Thus, it is important to perform exploration of this kind in a space that one understands well, or at the very least, to interpret the results cautiously and as provisional findings.

As we have noted repeatedly in this paper, the process is primarily exploratory. We believe it provides useful insight and generates hypotheses from existing data with relatively little effort and in a way that is complementary to other approaches. When implemented carefully, such results may be better than having no information, but all implications should be confirmed or checked with other methods.

There are many open questions about methods of this kind and the relationship between the stochastic operation of GAs (and other search algorithms) and the assumptions of the underlying data that come from HB and other preference estimation processes. For instance, it is conceivable that assumptions of HB (e.g., regarding distributions, error structure, and interactions) could interact with the search method in such a way that portions of the results are structured by the estimation process itself rather than by respondents' data. It may be very difficult either to confirm or dispute that possibility. We believe this is an important area for academic exploration; and in the intervening time, is yet another good reason to regard the output of this method as exploratory and generative in nature.

Another area for future research concerns the relationship among search methods and alternative ways to represent a portfolio. For instance, one might consider using search algorithms other than a GA (see Belloni et al, 2008) or portfolio methods derived from quantitative finance.

Computer Code

Computer code is available from the first author. It is free, open source, use-at-your-own-risk, and unwarrantied code provided solely for didactic and research purposes. The code implements the bootstrapped GA model for applications that provide standard HB utilities, such as the typical output of HB estimation in a Sawtooth Software CBC, ACBC, or similar conjoint analysis project. It is written in R and requires modest customization to the fitness function to implement pricing and attribute prohibitions or interactions (if any). We estimate that adaptation of the code to a given project should take approximately one day of analyst time, if the analyst has basic familiarity with the R language.

Options in the code include the ability to use either first choice or aggregate share of preference estimation; options to use HB draws instead of individual mean part worths; bootstrapping within the fitness function itself (e.g., across HB draws); and logit model exponent tuning. The default fitness function implements a share of preference model for portfolio fitness, but this could be adapted to implement fitness in terms of revenue, profit, directed competition, or other metrics.

An analyst may wish to explore optimization procedures other than a GA. In that case, it would be possible to use the provided framework to handle CA data and fitness assignment, but to replace the call to rgenoud that performs GA optimization and use another optimization routine instead. The primary requirement in that case would be to write an appropriate wrapper function that calls the optimization routine.

GA models are computationally intensive, and a large-scale project such as the one reported here may take days or weeks to run. The key aspects that increase run time are portfolio size, number of generations in the GA, and the GA population size; each of those produces a nearly linear increase in time. Computation time may be reduced either by simplistic brute force parallelization, such as running the model for different portfolio sizes on different computers; or by using a multicore workstation with parallelization options for the rgenoud library. The latter approach requires additional configuration of the R environment (cf. Mebane and Sekhon, 2009).

Conclusion

Search and optimization methods such as the GA approach offer analysts a way to mine existing data and derive insight along with potential opportunities for products. The method presented here offers a straightforward way to do this with HB data from discrete choice studies. The authors hope that the approach and the available code are useful to others, and we look forward to seeing future work from the choice modeling community.

Acknowledgements

The authors thank Ken Deal of McMaster University for insightful comments on the proposal, computer code, and presentation. Bryan Orme of Sawtooth Software commented on the approach and its relationship to other methods. Attendees of the 2010 Sawtooth Software Conference provided many interesting questions, comments, and related approaches; in particular, the audience observed how the fitness function could be scoped to different needs, and how the search algorithm itself could use a process other than a genetic algorithm. We have attempted to reflect those comments in this paper; oversights and errors are, of course, our own.

AUTHOR CONTACT

Questions about the computer code and other issues should be directed to Chris Chapman: chris.chapman@microsoft.com (primary), or cnchapman@msn.com (backup). Mailing address: Chris Chapman (cchap), 1 Microsoft Way, Redmond, WA 98052.

REFERENCES

Belloni, A., Freund, R.M, Selove, M., and Simester, D. (2008). Optimal Product Line Design: Efficient Methods and Comparisons. *Management Science* 54: 9, September 2008, pp. 1544-1552.

Chapman, C.N., Alford, J.L., and Love, E. (2009). Exploring the reliability and validity of conjoint analysis studies. Presented at Advanced Research Techniques Forum (A/R/T Forum), Whistler, BC, June 2009.

Chapman, C.N., Alford, J.L., Johnson, C., Lahav, M., and Weidemann, R. (2009). Comparing results of CBC and ACBC with real product selection. *Proceedings of the 2009 Sawtooth Software Conference*, Del Ray Beach, FL, March 2009.

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

Mebane, W.R., Jr., Sekhon, J.S. (2009). Genetic Optimization Using Derivatives: The rgenoud Package for R. [R package], http://sekhon. berkeley.edu/rgenoud.

R Development Core Team (2010). R: A language and environment for statistical computing [Computer software]. Version 2.11.0. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Sawtooth Software (2003). *Advanced Simulation Module for Product Optimization v1.5 Technical Paper*. Sequim, WA.

Sawtooth Software (2010a). CBC/HB 5.0 [Computer software.] Sequim, WA, 2010.

Sawtooth Software (2010b). SSI/Web 7.0 [Computer software.] Sequim, WA, 2010.

THE IMPACT OF COVARIATES ON HB ESTIMATES

KEITH SENTIS
VALERIE GELLER
PATHFINDER STRATEGIES[1]

INTRODUCTION

Since the ground-breaking work of Allenby and Ginter (1995) and Lenk, DeSarbo, Green & Young (1996), Hierarchical Bayesian (HB) analysis has been applied to a wide range of marketing problems with considerable success due to the ability of these methods to capture heterogeneity more effectively than do previous analytic frameworks such as aggregate logit. In the case of choice experiments, HB models enable estimates of individual level partworths that fit the heterogeneity in individual choice patterns (Huber, Arora & Johnson, 1998).

When market segmentation is on the agenda, this fitting of individual level partworths is an appreciated deliverable. Forty years after Smith (1956) defined market segmentation as making product decisions by studying and classifying the diversity of wants in the customers that define a market, HB provided a tool set that is highly appropriate for segmentation problems. HB analyses yield individual partworths by virtue of a two-level model in which the upper level model makes assumptions about the distributions of respondents' vectors of partworths and at the lower level, a logit model is assumed for each individual. In choice experiments, the amount of data available at the individual level is inadequate to fit individual models so HB uses information from the upper level model to assist with the fitting of the lower level model. More or less information is "borrowed" from the upper level model, depending on the extent to which a given respondent's choices are well estimated from his or her own data.

The assumptions that are made in the upper level model about the distribution of vectors of partworths come into focus, or indeed come under the microscope, when applying the HB toolset to segmentation problems. In the simplest case, the upper level model assumes that all respondents come from the same population and are drawn from a single multivariate normal distribution. The first four versions of Sawtooth Software's CBC-HB program incorporated this type of simple upper level model. This reliance on a single distribution in the upper level model that represents a single population of choice behaviour may present a conceptual stumbling block if the marketing problem concerns segmentation. In segmentation studies, the typical goal is to identify portions of the market that are different from one another and define these segments such that segment members are internally homogenous and externally heterogeneous. With this goal in mind, wheeling in an upper level model hampered with a single distribution is like trying to pin Hulk Hogan with one arm tied behind your back. How can a simple upper level model be appropriate given the constraint that members of all segments be drawn from the same multivariate normal distribution? Surely it would be better if the upper level model were sufficiently complex to allow members from each segment to be drawn from a separate upper level distribution.

It was this very question about what is in the upper level model when facing segmentation problems that prompted Sentis and Li (2001) to examine what happens when HB analyses with

[1] The authors thank Rich Johnson for many helpful comments.

simple upper level models are allowed to "borrow" information from more relevant subpopulations. In this initial work nine years ago, we first divided the sample into groups with presumably similar choice patterns and then estimated the HB partworths with a simple upper level model separately for each group. We attempted to improve predictive accuracy by having the analysis "borrow" information from an upper level model that was tailored to each specific segment rather than borrowing from the entire sample. Across seven commercial datasets, we compared the predictive accuracy of HB partworths derived from the entire sample to those derived from within *a priori* segments and from within two types of latent segments. To our surprise, we found that the 222 sets of partworths that were derived within segments yielded no improvement in predictive accuracy compared to 21 sets of partworths from the entire sample.

Reaction among colleagues to these 2001 results formed a continuum that was anchored at one end with comments like "this is so unintuitive you must have stuffed up something in your analyses" and at the other end by comments along the lines of "just what I expected because segments are not like individual cantaloupes but rather like slices of one single watermelon." We surmise that these reactions were grounded in antithetical worldviews about the nature of segments. These worldviews are fundamentally different in terms of how they represent the density distributions of respondents in space. One worldview posits that the density distribution of respondents looks like a bag of cantaloupes with each segment being represented by its own unique, albeit homogeneous cantaloupe. The other worldview argues that the density distribution looks more like a single watermelon and that segments are represented by slices of that common watermelon. Most of our colleagues subscribe to either the cantaloupe worldview or the watermelon-slice worldview of segmentation with very few people having a foot in both camps. In the aftermath of our 2001 work, our own worldview of segments had been severely challenged and we sought further data to crystallize our perspective on segments. We learned that Allenby, Arora and Ginter (1998) had examined three datasets looking for homogeneous segments (that is, cantaloupes) but did not find evidence of homogeneity of demand. More recently, Frazier, Jones and Patterson (2009) considered this issue within the context of MaxDiff problems. They examined three commercial datasets as well as three synthetic datasets and found no improvement in either model fit or parameter recovery when computing HB partworths within segments.

Given this body of empirical evidence that a custom-tailored upper level model does not improve predictive accuracy, we believe that the scales are tipped in favour of the worldview that segments are more akin to watermelon slices than to cantaloupes. Our own worldviews of segments notwithstanding, adherents to the cantaloupe model of segmentation have suggested alternative explanations for these results. One such explanation is that when there is sufficient information to fit the individual level model because each respondent has made an adequate number of choices, the upper level model has a limited effect on the HB estimates and therefore drawing from a single distribution rather than from segment-specific distributions does not hurt predictive accuracy. In discussions following the 2001 Sawtooth Conference, Joel Huber suggested that we try to "break" our findings by conducting the same analyses with fewer and fewer choice tasks. The idea was that by restricting the number of choice tasks included in the analyses and thereby reducing the amount of information available to the lower level model, HB would borrow more heavily from the upper level model and the improvement in predictive accuracy from within-segment analyses would be "revealed." After the conference, we conducted a series of HB analyses in which we reduced the number of choice tasks from 18 to 16

to 14 to 12 to 10 to 8 and reran the HBs on these increasingly sparse datasets. However, we were unable to "reveal" the effect that was expected by supporters of the cantaloupe theory of segmentation. Even with as few as eight choice tasks, within-segment analyses that enabled a segment-specific upper level model did not improve predictive accuracy.

Other "cantaloupe theory" supporters have suggested that within-segment computation of HB partworths fails to improve predictive accuracy because of the numerous additional parameters that are required when a new covariance matrix is estimated for each segment. This argument posits that the expected improvement in predictive accuracy is swamped by the overparameterisation involved when using a simple upper level model for each segment. An upper level model that incorporated separate distributions for each segment would be more appropriate and would avoid the overparameterisation problems inherent in the within-segment approach to HB analyses.

Previous research that incorporated covariates in HB models has examined the impact of covariates on predictive accuracy and reported only tiny improvements. Howell (2004) examined synthetic paired-comparison datasets and found that hit rates were improved by an average of 1.1 percent across two analyses. More recently, Orme and Howell (2009) conducted HB analyses of a commercial dataset and found that hit rates were improved by seven tenths of one percent and that holdout likelihoods were improved by five tenths of one percent when a covariate was incorporated in the HB model.

The fifth version of CBC-HB that incorporates covariates provides an easy-to-use tool for testing the overparameterisation explanation for our 2001 results. This new version of CBC-HB enables an upper level model of almost any complexity to be used in analyses of choice experiments. Armed with this more appropriate HB toolset, we felt cautiously optimistic about entering the ring against a few Hulk Hogan segmentation problems and set about asking the same question as in 2001, that is, in the context of segmentation problems, what should the upper level model look like? Specifically, is a custom-tailored upper level model better or should the KISS rule apply?

Approach

Our method for answering this question is straightforward. For five commercial datasets involving both services and FMCG with sample sizes ranging from 420 to 5,502 our analyses proceeded stepwise as follows:

Step 1: estimate HB partworths without a covariate

Step 2: estimate HB partworths using a covariate

Step 3: compare the quality of the two sets of partworths

Step 4: do this for several different covariates, one covariate at a time

We chose datasets in which a variety of potential covariates were available with particular emphasis of three classes of covariates: demographic variables, category behaviour variables and attitudinal variables.

Given that we have had more than two decades of experience in each of the relevant product categories for these datasets, we selected these particular covariates because, in our "expert

judgment", they would have an impact on the choice patterns of respondents. Our intention was to evaluate these three classes of covariates in terms of their differential impact of the quality of the HB partworths.

In the analyses that follow, we examined two different aspects of the quality of HB partworths:

- measures of model fit and of predictive accuracy
- measures of partworth variability

Measures of Model Fit and Predictive Accuracy

For each covariate within a given dataset, we computed these four measures of fit and predictive accuracy:

- **RLH** or "root likelihood" is a measure of goodness of fit of the model:
  - when the fit is perfect, RLH is one
  - if the model fits no better than chance, RLH is 1/k where k is the number of alternatives in the choice task
  - the higher the RLH, the better the fit
  - Example: if RLH = .800 for tasks with five choices, this means that the fit is four times higher than chance [.800/(1/5) = 4]
- **Hit Rate** is the percentage of actual choices on a holdout task that can be predicted using partworths that are estimated using the non-holdout tasks:
  - Hit Rate is a dichotomous measure of predictive accuracy:
  - for a given respondent on a given holdout task, the predicted choice is either correct and scored as a "hit" or it is incorrect and scored as a "miss"
  - the higher the Hit Rate, the better the predictive accuracy
- **Holdout Likelihood** is similar to Hit Rate in that it compares the predicted choice to the actual choice in hold out tasks:
  - unlike the Hite Rate, which is a dichotomous measure, Holdout Likelihood is a continuous measure of predictive accuracy
  - in Holdout Likelihood, the estimated partworths are used to calculate the share of preference for each of the alternatives in the holdout task for a given respondent
  - Holdout Likelihood is the average predicted share of the alternative that was chosen. Thus, for each holdout task, the Holdout Likelihood for a given respondent can vary from zero to one
  - the higher the Holdout Likelihood, the better the predictive accuracy
- **MAE** or "mean absolute error" is a measure of predictive accuracy of the aggregate shares in a fixed holdout task:
  - each respondent sees that same holdout task

- predicted share of preference for the alternatives in the holdout task is compared to actual share of preference
- absolute difference between predicted share and actual share (absolute error) is averaged across the alternatives
- the lower the MAE, the higher the predictive accuracy

Measures of Partworth Variability

For each covariate within a given dataset, we computed these two measures of partworth variability:

- importance spread
- standard deviation ratio

Importance Spread. The notion of "importance spread" was introduced by Orme and Howell (2009) as a metric for the extent to which the inclusion of covariates in HB analyses promote shrinkage to the segments' upper level model rather than shrinkage to a simple upper level model.

The calculation of "importance spread" is illustrated below for one dataset with a sample size of 836.

Step 1: calculate importance scores for each attribute:

Based on HB partworths estimated without covariate

Importance scores

RespNum	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6
1	66.26	266.29	15.92	140.2	68.58	42.75
2	4.38	358.99	5.66	149.21	19.26	62.5
…						
836	28.7	211.61	91.89	73.47	117.62	76.71

Based on HB partworths estimated with covariate

Importance scores

RespNum	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6
1	62.35	263.06	18.48	135.4	73.95	46.76
2	12.97	359.51	4.9	139.04	15.76	67.82
…						
836	35.01	207.08	85.97	76.63	116.39	78.92

Step 2: calculate mean percent importance score for respondents at each level of the covariate

Mean importance scores

	Mean Importance Scores					
	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6
State1	60.54	210.70	37.51	125.89	98.78	66.59
State2	65.57	201.44	40.26	125.46	99.68	67.59
State3	59.90	210.00	39.65	120.89	100.23	69.33
State4	67.99	203.91	41.40	119.47	100.50	66.72
State5	62.82	214.86	39.34	114.00	102.24	66.73

Mean importance scores

	Mean Importance Scores					
	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6
State1	60.25	212.39	36.27	126.56	99.06	65.47
State2	64.15	203.40	40.35	123.64	100.08	68.39
State3	58.05	205.66	40.89	119.91	101.09	74.40
State4	69.40	204.45	42.57	119.62	101.07	62.90
State5	62.18	213.21	39.02	112.82	104.80	67.97

Percent importance scores

	Percent Importance Scores						
	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6	Total
State1	10.09	35.12	6.25	20.98	16.46	11.10	100
State2	10.93	33.57	6.71	20.91	16.61	11.27	100
State3	9.98	35.00	6.61	20.15	16.71	11.55	100
State4	11.33	33.99	6.90	19.91	16.75	11.12	100
State5	10.47	35.81	6.56	19.00	17.04	11.12	100

Percent importance scores

	Percent Importance Scores						
	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6	Total
State1	10.04	35.41	6.04	21.09	16.51	10.91	100
State2	10.69	33.90	6.72	20.61	16.68	11.40	100
State3	9.67	34.28	6.81	19.99	16.85	12.40	100
State4	11.57	34.07	7.09	19.94	16.85	10.48	100
State5	10.36	35.53	6.50	18.81	17.47	11.33	100

Step 3: calculate spread as maximum percent importance score less minimum percent importance score

Minimum and maximum percent importance scores

	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6	
Maximum	11.33	35.81	6.90	20.98	17.04	11.55	
Minimum	9.98	33.57	6.25	19.00	16.46	11.10	
Spread	1.35	2.24	0.65	1.98	0.58	0.45	7.25

Minimum and maximum percent importance scores

	Attrib1	Attrib2	Attrib3	Attrib4	Attrib5	Attrib6	
Maximum	11.57	35.53	7.09	21.09	17.47	12.40	
Minimum	9.67	33.90	6.04	18.80	16.51	10.48	
Spread	1.90	1.63	1.05	2.29	0.96	1.92	9.75

Step 4: sum spreads across attributes and compute ratio

Importance spread ratio

	Sum of spreads	Spread Ratio
No covariate	7.25	1.34
Covariate	9.75	

Standard Deviation Ratio. This measure of partworth variability is perhaps more intuitive insofar as it is based on the familiar standard deviation statistic. The steps in calculating this metric are illustrated with the same dataset as we used to illustrate Importance Spread.

Step 1: calculate standard deviation for each attribute's partworths.

Based on HB partworths estimated without covariate

	Standard Deviations				
	PW1	PW2	PW3	...	PW30
836 Resps	0.59	0.50	0.76		0.56

Based on HB partworths estimated with covariate

	Standard Deviations				
	PW1	PW2	PW3	...	PW30
836 Resps	0.61	0.51	0.77		0.57

Step2: for each attribute, calculate ratio of standard deviation of partworths with covariate to standard deviation of partworths without covariate

	Ratio of Standard Deviations				
	PW1	PW2	PW3	...	PW30
836 Resps	1.04	1.02	1.02		1.01

Step 3: calculate mean of these ratios across entire set of partworths

Mean Ratio	1.02

RESULTS

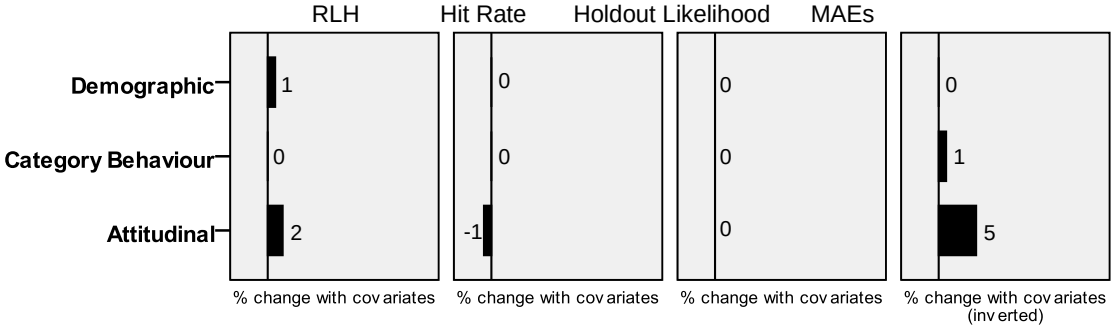
In the graphs that follow, we summarise the results of our analyses of five commercial datasets:

- n = 8,445
- 40 attributes
- 189 parameters
- 43 covariates (19 demographic, 12 category behaviour, 12 attitudinal)
- 4,900,000 iterations

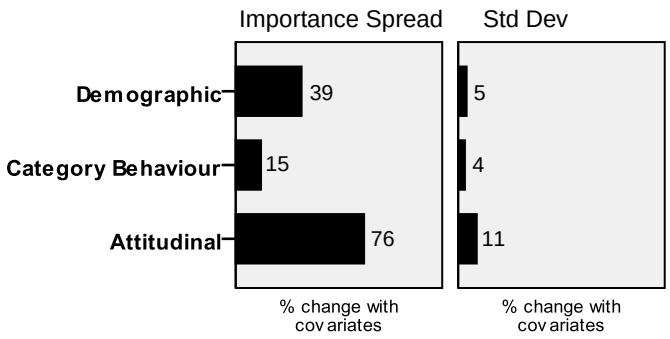
These 43 covariates were selected based on expectations that they would effectively delineate segments of respondents with different choice patterns. To confirm our expectations, we analysed the alpha draws and found that with very few exceptions, the draws for several partworths in each of the covariate models were consistently positive or negative.

Measures of Model Fit and Predictive Accuracy

The four graphs below show the percentage improvement in the measures of fit and predictive accuracy that were obtained by the inclusion of a covariate in the HB model. The four measures are broken out by the three classes of covariates. It is clear that these measures are essentially unchanged by including any of the three classes of covariates. Given the complete lack of impact, we failed in our quest to provide substantive advice about the three classes of covariates.



The two graphs below show the percentage improvement in the measures of partworth variability. The inclusion of covariates when estimating HB partworths resulted in increased partworth variability.



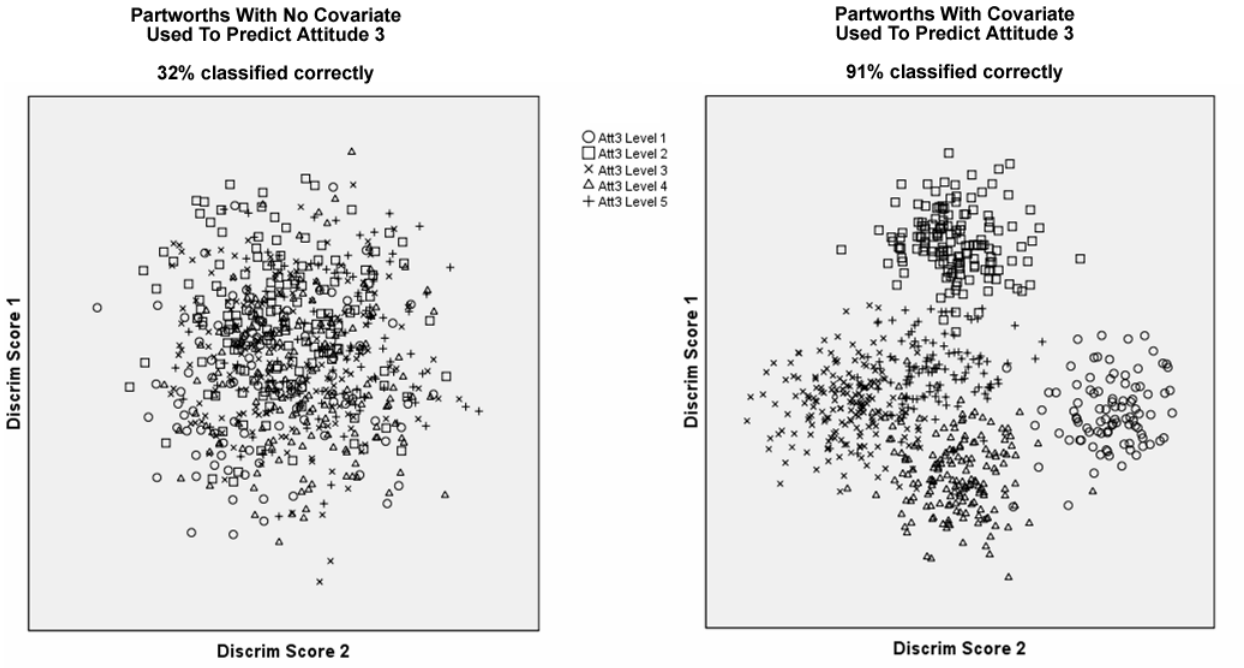
We had originally planned to complete these analyses on ten commercial datasets. However, when the results from our analyses of the first five datasets proved to be the same, we turned our attention to possible explanations for why predictive accuracy is not improved by the inclusion of covariates.

Our first hypothesis was that the partworth estimated using covariates are not all that different from those estimated without covariates. To test this hypothesis, we ran MANOVAs on the HB partworths computed with and without the different covariates used in our third dataset, which had these characteristics: n=714, 10 attributes, alternative specific design, 60 parameters, 7 concepts per task with dual none, 14 random and 2 fixed holdout tasks, 14 covariate variables.

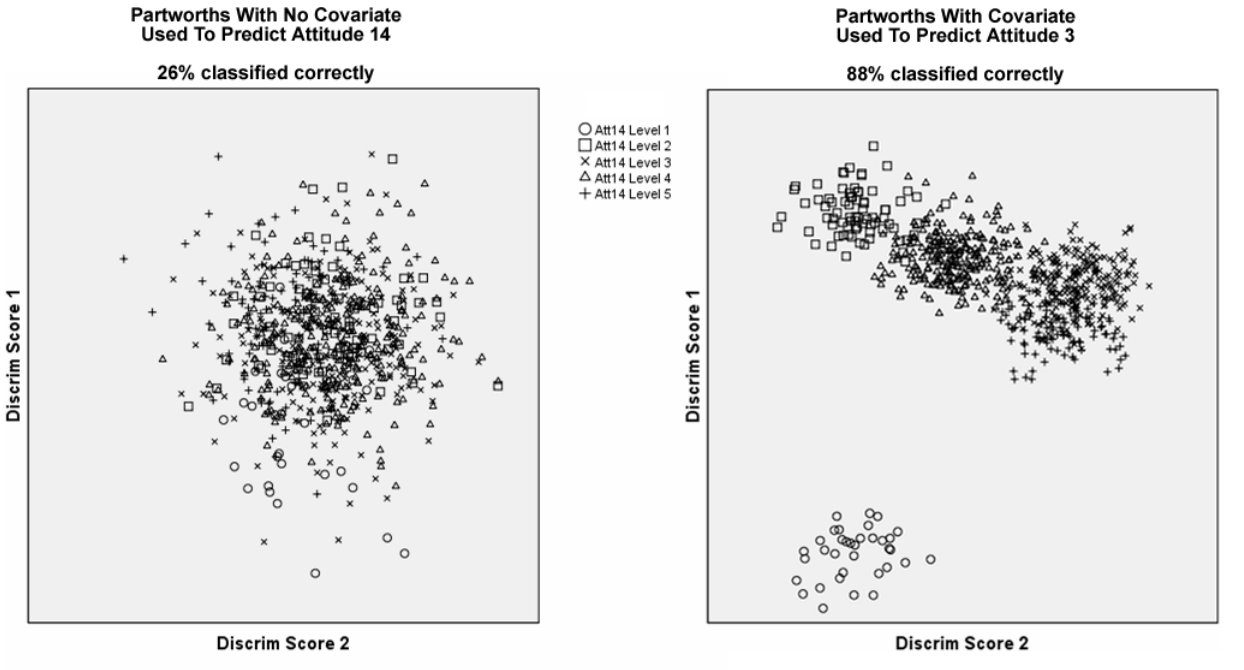
These MANOVAs indicate that the main effect (covariate or no covariate) and the interactions (covariate/no covariate by partworth) are highly significant. Thus, in statistical terms, the partworths estimated with covariates are indeed different than those estimated without covariates.

Using this same dataset, we conducted discriminant analyses to provide a graphical demonstration of these differences in the partworths that are estimated with and without covariates. First, we used the partworths estimated without covariates to predict segment membership for a five level attitudinal covariate. This first discrim resulted in only 32% correct classification of segment membership. The scatter plot on the left illustrates this analysis.

Next, we used the partworths that were estimated with this attitudinal covariate to predict segment membership. This second discrim achieved 91% accuracy in recovering segment membership as illustrated in the scatterplot on the right. These two scatter plots illustrate the extent to which partworths estimated with a five-level attitudinal covariate are specific to the five levels of the covariate.



We replicated the same set of analyses on a different five-level covariate from the third dataset with essentially the same results. The two corresponding scatterplots from this replication are shown here.



The MANOVAs and the two sets of discrims were sufficient evidence for us to reject our first hypothesis that partworths estimated with covariates are not different from those estimated without covariates. Our analyses indicate that HB partworths estimated with covariates are

different from those HB partworths estimated without covariates and they are specific to the levels of the covariate.

Having dismissed our first hypothesis about why the inclusion of covariates does not improve predictive accuracy, we came to a second hypothesis. Namely that HB partworths, which are estimated using covariates, are overfitting the choice data. That is, these HB partworths with covariates are fitting "noise" rather than "signal" in the choice data of the different covariate segments.

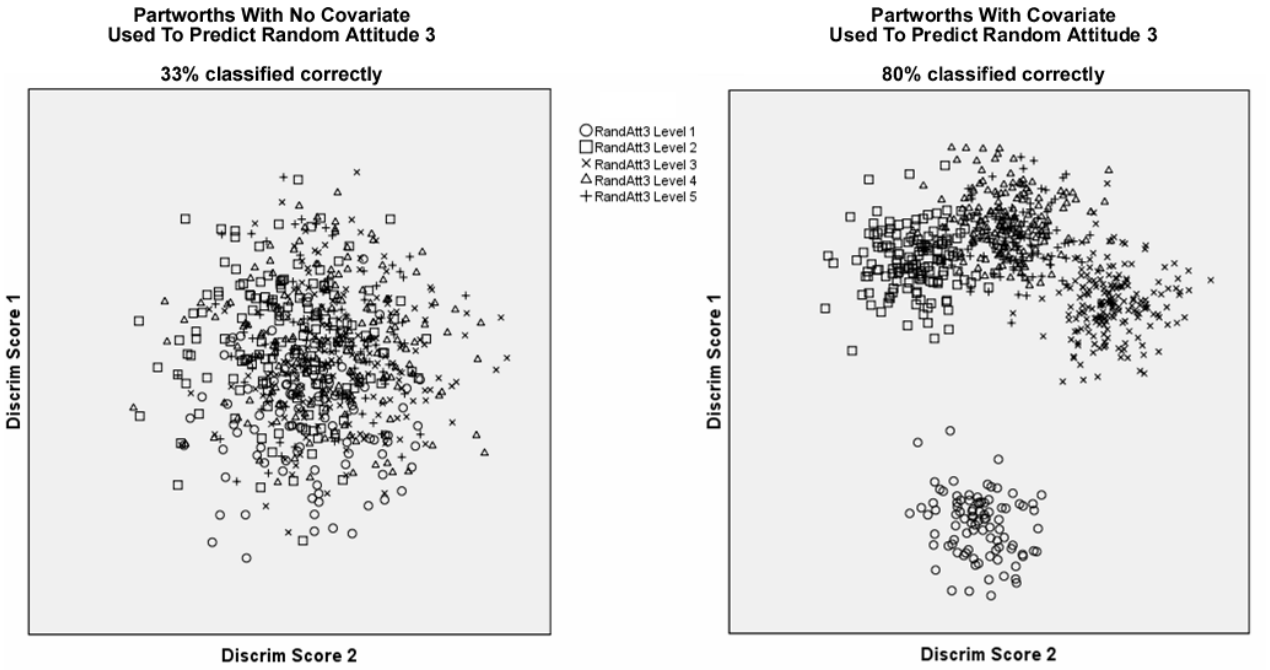
We conducted a simple experiment to test this hypothesis:

- we randomly shuffled the values of the covariate across respondents such that no respondent had the same level of covariate before and after the shuffling. As shown in the table below, while the segment sizes are the same for the actual covariate data and the random covariate data, the empty diagonal indicates that none of the respondents was shuffled into the same segment

		Actual Covariate Level					
		1	2	3	4	5	Total
Random Covariate Level	1	0	16	18	22	36	92
	2	35	0	47	44	22	148
	3	19	79	0	50	38	186
	4	19	31	96	0	23	169
	5	19	22	25	53	0	119
	Total	92	148	186	169	119	714

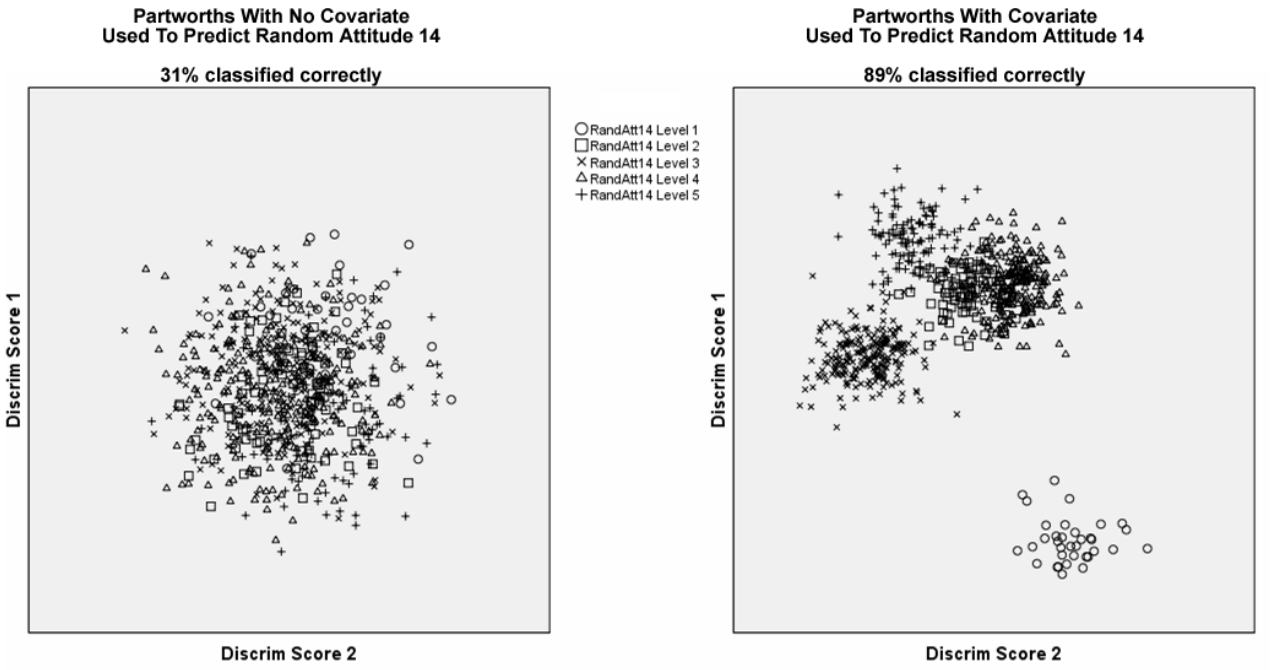
- by shuffling the covariate values, we completely eliminated the "signal" that the covariate was providing in the HB analysis of the choice data
- we estimated HB partworths using the random covariate and used another set of discrims to compare those partworths to the HB partworths that were estimated without a covariate

The discrim on the random covariate using partworths that were estimated without a covariate correctly classified segments members with 33% accuracy. However, the discrim based on the partworths estimated using the random covariate was able to correctly classify 80% of the randomly shuffled segment members. The scatterplots below from these discrims on the random covariate illustrate the extent to which the HBs estimated with a random covariate are able to accurately recover the randomly assigned covariate. That is, the partworths estimated with a the random covariate are accurate in predicting "randomness".



For completeness, we replicated this analysis by randomising the other attitudinal covariate, then estimating partworths using this random covariate and running the same set of discrims. The discrim using partworths estimated without the random covariate was able to recover the random covariate membership with 31% accuracy. The discrim based on partworths that were estimated using the random covariate achieved an 89% correct classification of the randomly assigned segment members. These discrims are illustrated in the two graphs below.

Once again, we were able to accurately recover the random segment membership using HB partworths.

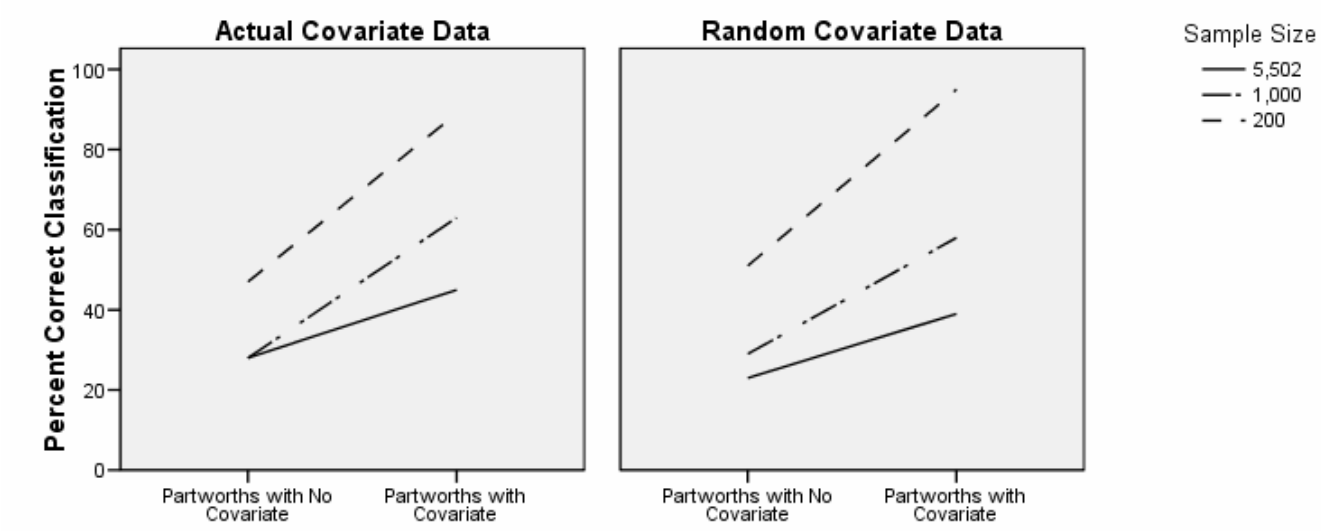


To further explore this explanation, we conducted these same analyses on our fifth dataset that has 5,502 respondents. As before, we used a five-level covariate and shuffled respondents on this covariate, ensuring that the segment sizes remained the same but that no respondents were in the same segment. We ran out HB partworths using this random covariate followed by the two discrims as in the previous analyses. However, for this dataset, we also selected random subsamples of 1,000 and 200 respondents. We then ran HB analyses with no covariate and with the actual and randomised covariates on these smaller subsamples.

If overfitting is responsible for the ability of HB partworths to recover randomly assigned segment membership, then this effect ought to increase as we reduce the sample size. If overfitting is not the culprit, then there should be no systematic effect when we reduce the sample size.

The graphs below shows quite clearly that as sample size is decreased, the ability of HB partworths to recover segment membership increases. This increase occurs in roughly equal measures regardless of whether the segment membership is actual or random. Thus, these analyses on the fifth dataset provide strong (albeit indirect) evidence that HB analyses with covariates are overfitting the choice data.

Correct Classification of Segment Membership
For Actual and Random Segments



Finally, we draw attention to the slopes of the lines in the graph of random covariate data. Compared to the slope when n=5,502, there is an increase in the slope of the line even with an n of 1,000, which is a sample size that is considered "generous" in HB circles. Thus, our indirect measure of overfitting suggests that this issue may be prevalent in datasets that hitherto were thought to be immune.

CONCLUSIONS

These analyses have shown that including covariates in HB analyses of choice experiments does not improve predictive accuracy. (We note that Kurz and Binner (2010) came to the same conclusion in their paper at this year's conference.)

We tested two hypotheses about why covariates do not increase predictive accuracy in HB analyses. The first hypothesis was that the HB estimates with and without covariates are not different but this hypothesis was rejected.

Our second hypothesis focused on overfitting. We developed an indirect measure of the degree of overfitting in HB models with covariates. Using this indirect measure, we found strong evidence that overfitting is an underlying cause for the failure of covariates to improve the predictive accuracy of HB models.

While considerable work remains to be done using synthetic data to further examine our findings, our work suggests the following:

- claims that overfitting is not a problem with HB, even with large samples, may be exaggerated
- caution is the order of the day when including covariates in HB analyses
- advocates of the cantaloupe model of segments are standing on shaky ground

REFERENCES

Allenby, G. M.; Arora, N. and Ginter J. L (1998) "On the heterogeneity of demand." Journal of Marketing Research 35, 384-389.

Allenby, G. M. and Ginter J. L. (1995) "Using Extremes to Design Products and Segment Markets." Journal of Marketing Research, 32, 392-403.

Frazier, C., Jones, U. and Patterson, M. (2009) "Estimating MaxDiff Utilities: Dealing with Respondent Heterogeneity." Sawtooth Software Conference Proceedings, Sequim, WA, pp 285-290.

Howell, Well (2004) "Reducing Respondent Burden in Ranking Tasks: Hierarchical Bayesian Analysis of Pairwise Comparisons with Covariates." Joint Statistical Meeting - Section on Statistics and Marketing, Activity Number 192, Toronto.

Huber, J., Arora, N. and Johnson, R. (1998) "Capturing Heterogeneity in Consumer Choices." ART Forum, American Marketing Association.

Kurz, P and Binner, S. (2010) "Added Value through Covariates in HB Modeling. " Sawtooth Software Conference Proceedings, Sequim, WA, pp 269-282.

Lenk, P. J., DeSarbo, W. S., Green P. E. and Young, M. R. (1996) "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs." Marketing Science, 15, 173-191.

Sentis, Keith and Lihua Li (2001) "One Size Fits All or Custom Tailored: Which HB Fits Better?" Sawtooth Software Conference Proceedings, Sequim, WA, pp 167-177.

Smith, W. (1956) "Product Differentiation and Market Segmentation as Alternative Marketing Strategies." Journal of Marketing, 21, 3-8.

ADDED VALUE THROUGH COVARIATES IN HB MODELING?

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

SUMMARY

We re-analyzed ten CBC data sets, comparing the use of covariates in HB to standard HB runs that assume single multivariate-normal populations. With HB using covariates, respondents are not shrunk toward one common distribution. Instead, part-worths of respondents with different characteristics have different multivariate-normal distributions. From a theoretical point of view this would seem more appropriate. We tried to find out whether in practice the use of covariates offers gains in predictive validity with respect to holdout choices and real market data.

We found few significant gains in predictive validity when including covariates - no matter whether the covariates were demographics, cluster segments, or segments based on past behavior and purchase intention questions.

As a second question we analyzed whether covariates could stabilize the estimates when there are reduced numbers of respondents and choice tasks. But reducing the amount of data for HB estimation in either way did not affect outcomes much. Also when analyzing within different segments we didn't find meaningful differences in outcome with and without covariates.

Our last section deals with whether covariates could improve matters when using proportional sampling within small segments of the population. We found that covariates couldn't resolve previously identified problems with proportional sample structure. In the small segments, there were small improvements with covariates, but results were far inferior to those of proportional sampling.

INTRODUCTION TO COVARIATES

The hierarchical Bayes (HB) model is called hierarchical because it models respondent preferences as functions of an upper-level (averaged across sample) model and a lower level (individual respondent-level) model. At the lower individual-level, the respondent is assumed to choose concepts by maximizing the sums of part-worth utilities as specified in the multinomial logit model.

In the standard HB approach, the upper level model assumes that respondents are drawn from a single multivariate normal distribution, with part-worths (ßi) distributed with means α and covariance matrix D, ßi ~ Normal(α, D), where i indicates the single respondent. In HB applications, the upper-level model plays the role of a prior distribution when estimating each respondent's part-worths, and the lower-level model provides the likelihood for the estimation. Because it leverages information from the upper-level population parameters α and D, HB is able to estimate relatively stable part-worths for each individual, even when the data set provides only relatively sparse information.

The assumption of a single multivariate normal population is troublesome to some researchers, who consider most markets to be composed of distinct segments. Many researchers have considered ways to modify the upper-level model so as to be more compatible with the assumption of discrete segments. We review some of those attempts before looking at covariates.

In standard HB approaches a simple assumption is used: respondents are drawn from a single population of normally distributed part-worths. While this assumption may seem to be very simple from a theoretical point of view, it performs well in most of our studies. The single-normal-population assumption is often an influencing factor only at the start-up of the estimation and does not affect the final part-worth estimates to a large extent. Especially, it does not constrain the final part-worths to be normally distributed. HB results represent a combination of the upper- and lower-level models for each individual. If enough information is available at the individual level, the resulting part-worth utilities don't show much influence of the upper-level model compared to the impact of the lower-level model. Unfortunately the influence of the upper-level model results in some Bayesian shrinkage toward the population mean value, which tends to smooth the distribution, with a tendency toward the normal distribution (especially if information is sparse for an individual). But again, if a substantial number of choice tasks are available in relation to the number of parameters to be estimated, the Bayesian shrinkage is usually small and therefore doesn't affect the result very much.

In some instances, practitioners see problems with the assumption of a single normal population:

- The assumption that respondents are drawn from a single normal population seems for many practitioners and clients unrealistic; they assume more complicated functions.
- In segmentation studies, practitioners have expressed concern that distances between segment means are shrunk because HB tends to shrink individual estimates towards the population mean.
- In situations in which a segment of respondents (with different preference structures) is oversampled, this Bayesian shrinkage can bias the estimates for the segment means as well as the overall population means – especially for the proportional part of the sample (Fuchs, 2007).

In recent years, researchers have proposed ways of solving this problem. One is the idea from Sentis and Li (2001) of estimating for segments separately, to avoid the shrinkage problem. The idea is based on the fact that, if we use a different population mean value for each segment, shrinkage to the overall population mean would no longer occur. Sentis and Li studied seven actual CBC data sets, systematically excluding some of the tasks to serve as holdouts for internal validation. They estimated the utilities in four ways: first by using the entire sample within the same HB estimation routine (one population mean value); second by segmenting respondents according to industry sectors and estimating HB utilities within each segment (mean value for each segment); third segmenting respondents using a K-means clustering procedure based on first stage HB utilities, and then re-estimating within each segment using HB; and fourth by segmenting respondents using Latent Class and then estimating HB utilities within each segment. They found that none of their attempts to improve results by fitting subgroups separately improved predictions of holdouts.

Howell, (2007) proposed respondent weighting in HB as a solution to the disproportional sampling problems outlined in the Fuchs (2007) paper. He investigated the severity of the problem, and used simulated data to demonstrate that when subgroups are dramatically oversampled, it causes the means of smaller groups to shrink disproportionately toward the larger groups. This could bias the sample means for the proportional (under-represented) groups, and violates the accuracy of preference share simulations. Howell shows that much of the problem is due to diverging scale factors between smaller and larger subgroups. The scale for the oversampled (disproportional) groups is expanded, leading to stronger pull on the overall sample mean. The article shows with artificial data that normalizing the scale post hoc can largely control this issue. Also it concludes that implementing a simple weighting algorithm within HB (computing a weighted alpha vector) can potentially improve matters further when there are extreme differences in sample sizes between subgroups. Practitioners often deal with the problem by using disproportional sampling and estimating the groups separately for each of the small segments to avoid Bayesian shrinkage.

Other approaches to solve the problem employ multiple upper-level distributions. It is apparent that continuous heterogeneity (normal mixture) alone is superior to discrete heterogeneity (Latent Class), up through a fairly large number of segments (Rossi, Allenby & McCulloch 2005); also, a correlated (random) coefficients specification for the normal mixture is superior to an uncorrelated one; and more than one segment can be used in the normal mixture model. But for multiple mixtures of upper-level models, more complex mathematical functions and estimation procedures are needed as well as a lot of prior knowledge about the data structure to gain the right multivariate-normals. Allenby & McCulloch (2005) found that extending HB to accommodate multiple distributions leads to only minimal gains in predictive accuracy. From a practitioners' point of view we can solve this problems by estimating the relevant segments separately if we know the data structure upfront.

Advanced HB practitioners have recommended that in many cases "well-chosen" covariates could provide additional information and therefore improve parameter estimates and preference share predictions (Lenk, et.al. 1996). Covariates could be seen as another term using additional independent variables that may affect part-worths. Often, we think of covariates such as demographics like gender, age, income, company size, geographic location, etc. Unfortunately, these variables often have only low correlations with the preference structure of our choice context. The most useful covariates bring exogenous information (additional information which is not already available in the choice tasks) to the model to improve the estimates of part-worth and improve preference share predictions.

More formally, instead of assuming respondents are drawn from one normal distribution with mean vector α and covariance matrix D, an HB model which uses covariates in the upper-level model assumes that respondent part-worths are related to the covariates through a multiple regression model:

$$\beta_i = \Theta' z_i + \varepsilon_i \text{ where } \varepsilon_i \sim \text{Normal}(0,D)$$

where Θ is a q by b matrix of regression parameters, z_i is a q vector of covariates, and ε_i is a b vector of random error terms. The part-worths are now drawn from a normal distribution with mean values $\Theta' z_i$, , different for each respondent. No longer shrinking the individual estimates to a single population mean α, this method shrinks them to the conditional mean Θz_i given the

subject's covariates. With this solution the multiple regression upper-level model can use observed, segment basis variables (e.g. Country, Car Segment, Distribution Channel, etc.) to improve the estimation of the part-worths and may increase the distinction between segments in the data set.

In the standard HB model with the single normal population mean value, there are b + [b(b+1)]/2 parameters to be estimated in the upper-level model, where b is the number of part-worths for each individual respondent. When including covariates in the upper-level model, there are bq + [b(b+1)]/2 parameters to be estimated in the upper model, where b is again the number of part-worths and q is the number of parameters introduced by the covariates. If Country was the covariate, consisting of China, Russia, Italy, UK, US and Germany, q would equal 6. Including covariates in the upper-level model doesn't alter the number of parameters estimated for covariance matrix D. Using covariates is more parsimonious than separating the sample by country and running standard HB with a different covariance matrix for each of the six separate samples. In our data set number 6, the vector of sample means plus the covariance matrix require 49 + [49(49+1)]/2 = 1274 parameters to be estimated for the upper model in each sample, for a total of 6 * 1274 = 7644 parameters if samples are estimated separately. Estimating as a single run with a dummy-coded covariate for country requires only 1519 parameters in the upper-level model, a very substantial saving over the number required when running standard HB within the separate segment samples. So, using covariates requires estimating many more parameters than with the single-normal "standard" model, but many less than when estimation is done separately for subgroups. Using covariates also takes more computer time than the standard single normal approach, but less than making separate runs for each segment. For more information about the mathematics of the HB model with covariates introduced, please see Orme & Howell (2009).

Description of Analyzed Studies

For the purpose of this paper 10 commercial studies with a total of approximately 30,000 interviews covering nearly all industries and topics were analyzed. These studies cover B2B as well as B2C markets and were conducted in almost all parts of the world. All of these studies were carefully designed and used Sawtooth Software. As to good research practice and to ensure valid results, the sample structures of these studies were disproportional, ensuring sufficient sample size for all segment cells.

	Industry	Target Group	N =	Interview Delivery	# Coice Tasks	# Est. Parameter	Concepts/ Task	Conloint Method	Type of Natural Segments
Study 1	Tyres	B2C	189	CAWI	16	19	4	STD CBC	4 customer types
Study 2	DIY	B2C	500	CLT/CAPI	12	8	5	CBC ASD	5 distribution channels
Study 3	Adhesive	B2C/B2B	888	CAWI	8	40	16	CBC ASD	2 customer types
Study 4	Fuel Cells	B2C	926	CAWI	16	69	4	CBC ASD	3 countries
Study 5	Adhesive	B2C / B2B	600	CAPI	16	13	5	CBC ASD	5 customer types
Study 6	Automotive	B2C	8900	CAPI	15	49	5	STD CBC	6 countries x 12 car segments
Study 7	Automotive	B2C	8400	CAPI	14	86	3	CBC ASD	12 car segments
Study 8	Automotive	B2C	9200	CAPI	19	21	6	CBC ASD	6 countries
Study 9	Technology	B2C	3000	CAWI	15	74	8	CBC ASD	4 countries x 3 customer types
Study 10	FMCG	B2B	750	CAPI	18	84	12	CBC ASD	5 distribution channels

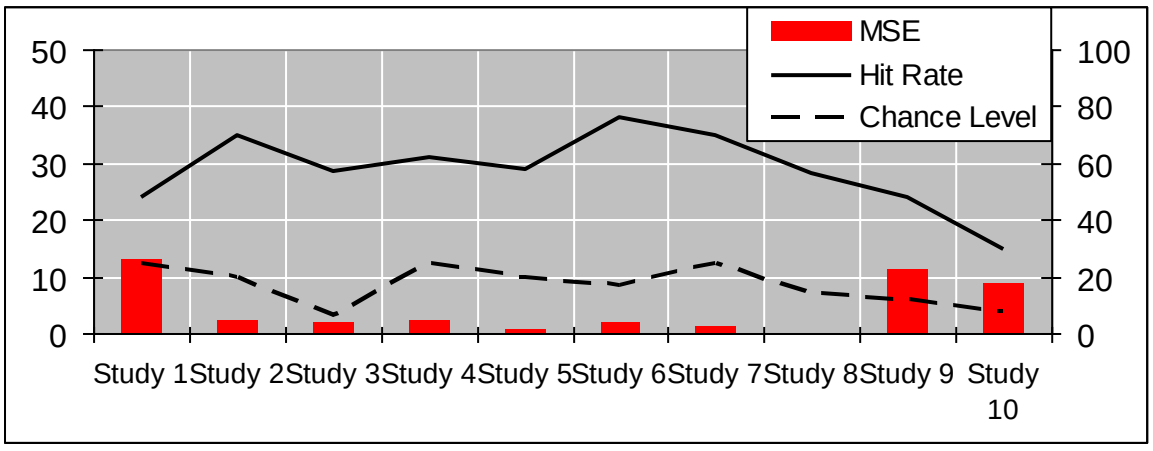
Table 1: Overview of studies analyzed

The 10 studies have been selected as they represent different scenarios from practical modeling work with conjoint analysis.

Study 1	Tires	Study on motorcycle tires in one country. Due to the different bike types in the study there is quite a lot of heterogeneity among the respondents
Study 2	DIY	Price Conjoint which was conducted in four different distribution channels which have different competitive environments
Study 3	Adhesive	This study analyzed the impact of branding on price elasticity in both, professional and private end user markets
Study 4	Fuel Cells	Study about energy supply in RVs. Very complex model with different product alternatives and 69 parameters to estimate. Difficult to recruit target group led to comparable small sample size in each of the three countries analyzed
Study 5	Adhesive	Brand/price conjoint in market with highly fragmented customer segments (some B2B, some B2C)
Study 6	Automotive	Automotive Study with different car features 6 countries 12 car segments from small mini cooper up to a lager limousine and suv's – sportscars face2face computer assisted 15 choice tasks 49 parameters Number of respondents compared to number of parameters should result in very stable estimates.
Study 7	Automotive	Automotive newer concepts of engines hybrid, active hybrid gas engines 12 segments one country more parameters asd model
Study 8	Automotive	Tires 9000 6 countries small number of parameters less concepts
Study 9	Technology	Technology 4 countries 3 typs of customers flatscreens 3000 online 74 parameter
Study 10	FMCG	10 fmcg b2b manufacturer of chips respondents retailers capi 84 parameters 5 channels hyper markets to groceries

Table 2: Description of studies analyzed

We examined both the hit rate for predicting the holdout choice, as well as the mean square error (MSE) of the base case simulation against the holdout task results. When looking at this performance measure standard HB (in the following labeled as HB STD) showed rather satisfying results in regard to these two measures:



Graphic 1: MSE and Hit of standard HB in ten studies

Types of Covariates used for Analysis

For the systematic analysis of the ten commercial studies, different types of Covariates were defined:

Type 1: Membership in Natural Segment

Demographic or product specific segments (categorical variables) were used as covariates. These were for example countries, customer groups, product segments or distribution channels. For further analysis in this paper we labeled models with this type of covariate as HB COV-N (HB with covariate based of natural segment), models with independent HB estimations for every single natural segment were labeled as INDV HB.

Type 2: Membership in Segments

For this group of covariates we used Latent Class or benefit segments (categorical or dummy coded). The benefit segments were derived by cluster analysis of individual utilities. Covariates based on Latent Class segments are called HB COV-L, those based on benefit segments HB-COV-U.

Type 3: Added Data

For a limited number of studies additional data was available and used as Covariates. Such added data included purchase intention (e.g. stated budget for new vehicle) or past behavior (e.g. purchase price of last vehicle)

All ten studies were analyzed with the these types of estimation models

HB STD	Standard HB for the whole study sample
HB COV-N	HB with covariate (defined by natural segment)
INDV HB	Independent HB Estimations per natural segment
LC	Latent Class (Sawtooth Software)
HB COV-L	HB with covariate (defined by LC segments)
HB COV-U	HB with covariate (defined by STD HB utility cluster)

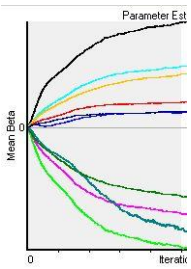
All estimations were performed with Sawtooth CBC/HB (v5.2.2) using standard settings (20,000 iterations, prior variance 2, degrees of freedom 5.

Observations during the Estimations

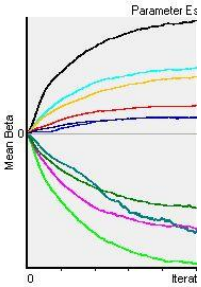
We could observe during the estimations that the models showed in first 1,000 - 10,000 draws a different behavior in convergence, while in the end the models converged to the same parameters than without covariates. We assume that this is caused by an influence of the upper-levels model when using covariates.

Following convergence plots demonstrate the slightly different behavior (shapes) at the beginning while finally converging towards the same parameters.

Without Covariates:

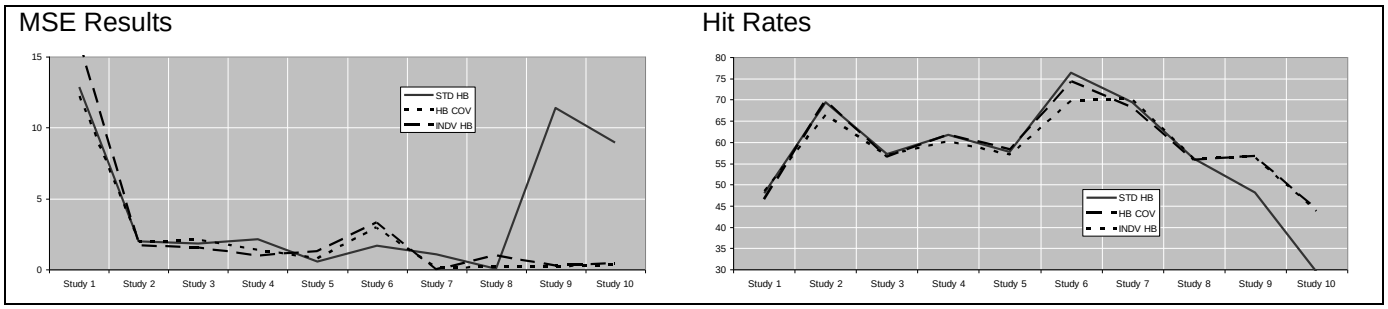


With Covariates:



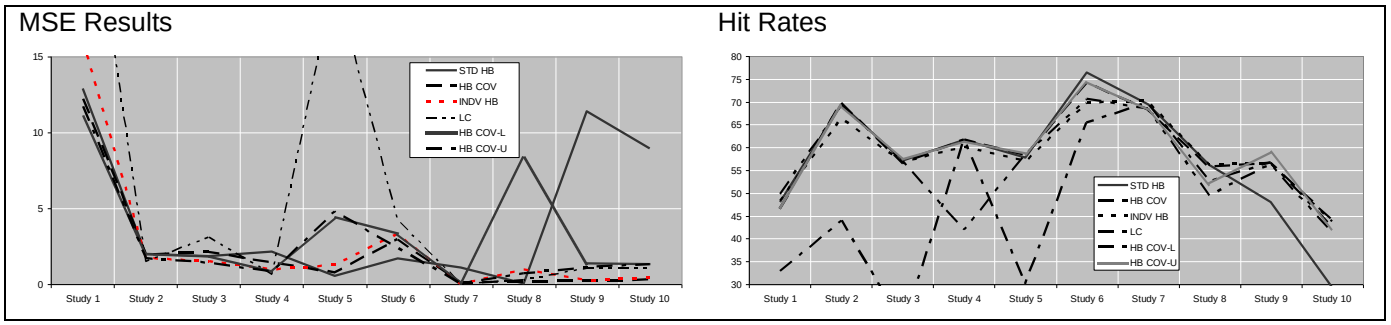
RESULTS OF DIFFERENT ESTIMATIONS

Only in two of ten studies HB COV showed significantly better MSE results than STD HB. The reason for this might be the relatively large number of parameters and small samples for those two studies (study nine and ten). However in these two studies HB COV was not better than INDV HB. Looking at the Hit Rates we observed the same results.



Graphic 2: MSE and Hit Rates of different simulation models

On the other hand there was no real champion among the alternative estimation methods based on segment membership or Latent Class:



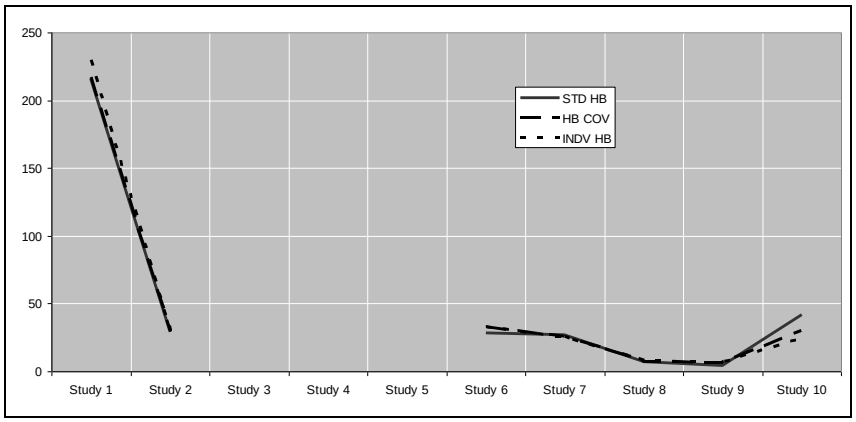
Graphic 3: MSE and Hit Rates of different simulation models

Especially the results of LC showed a controversial picture: In study 5 LC led to the worst MSE result while the Hit Rates (based on cluster members averages) scored best.

Overall there was no significant improvement in most of the studies through usage of covariates (either with natural segments or LC or Utility cluster based). Also, other estimation methods like LC or single HB segment estimation did not exceed the results achieved with standard HB in a significant way.

Holdout task results are mostly used for measurement of MSE and individual Hit Rates. As we had real market data for 7 of the 10 studies we used this calculate the MSE of the non-calibrated simulations (no correction for distribution and other external effects) as the ultimate proof of validity.

With exception of study one, which has a small and fragmented sample as well as a simplified attribute/level model, there were neither real differences between real market data and data from our studies nor significant differences between STD HB, HB COV and INDV HB as the graphic below shows:



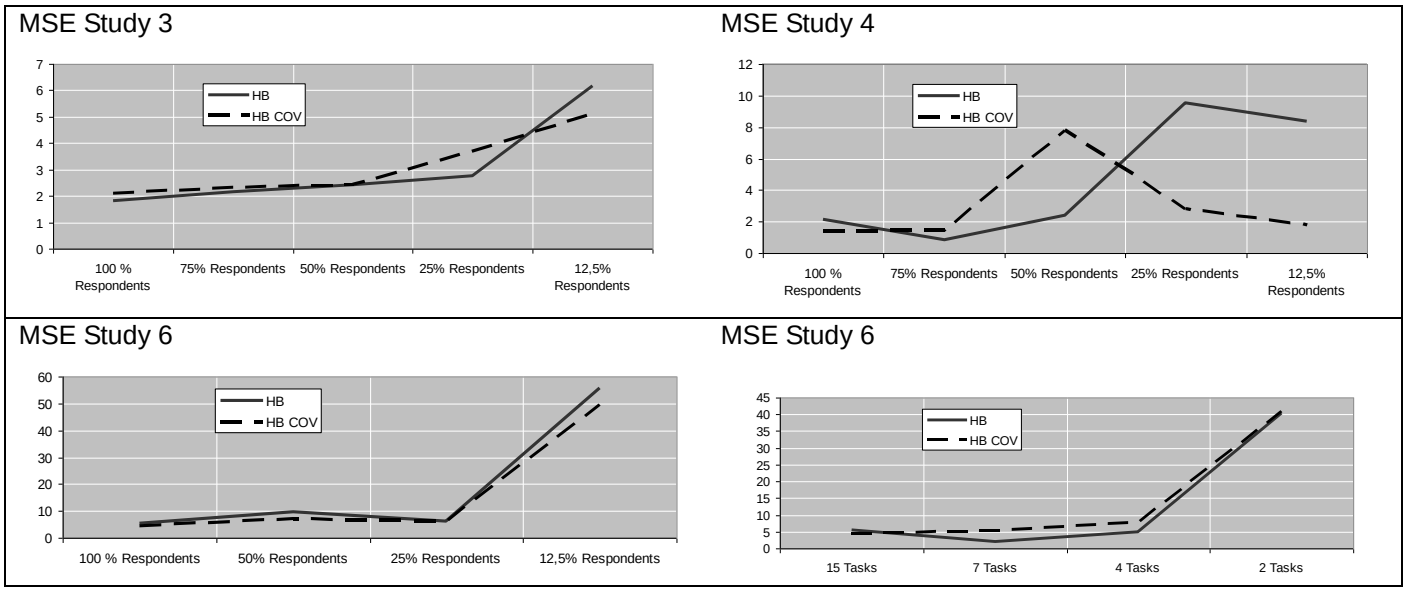
Graphic 4: MSE of not calibrated simulations against real Market Data

Our first conclusion is that studies which are set up correctly and have large enough sample size in all subgroups don't show better estimates when using covariates.

ESTIMATIONS WITH WEAKENED DATA

Based on our experience with the different estimation models and in order to simulate sparse data sets or poorly designed studies, we analyzed the effect of covariates on weakened data. In 3 of the ten studies we reduced the number of respondents stepwise randomly from 100% to 25%. The next try was to reduce number of tasks stepwise from 15 to 2 tasks by deleting the later tasks from the interview process.

To our surprise the reduction of "some amount" of information had no significant effect on the estimations. In two studies the MSE results remained on the same level even with only 50% of respondents, or as in study 6, with 25%.



Graphic 5: Simulation results with stepwise weakened data

Only in study 4 HB COV helped after reducing the sample by 50% or more. Otherwise there was no big difference between the HB with and without covariates. Furthermore once the information lack became too large (e.g 15% of sample) the error increased dramatically. However, again the Covariate was also not able to improve the simulation results.

The second conclusion is therefore that covariates are not a "first aid kit" for badly designed studies. Even though information could be reduced to some extent without damage, sample size must be retained to ensure representativeness.

As all of the previous analyses were based on the accuracy of prediction for the markets covered by the ten different studies, it was necessary to investigate how covariates could impact on the simulation of single segments of a study respectively of a market. For this purpose we selected two of the 10 studies and ran MSE and Hit Rate analysis within the natural market segments. As graphic 6 shows, there were no significant differences between the different estimation models. In study 4 there were three segments, and in each segment a different estimation model performed best. In the six segments of study 5 there was also no clear winner:

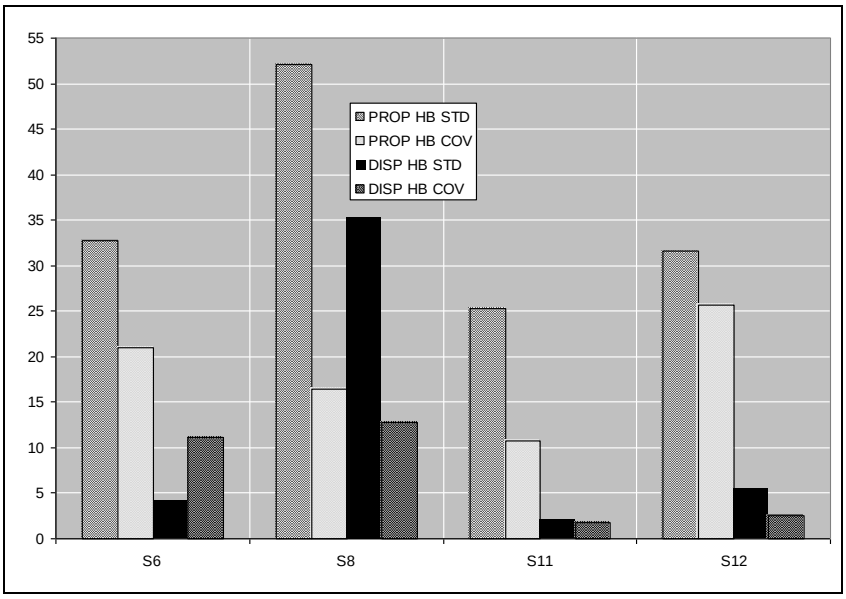
Graphic 6: Simulation results within market segments

As there was no significant effect in within-segment estimation between different variants of HB estimations to be observed, the third conclusion is that with disproportional sample structure (same/sufficient sample size for segment cells as there were in all ten studies) no improvement can be achieved by using Covariates.

ESTIMATIONS WITH PROPORTIONAL SEGMENTS

Study 6 had a quite large sample size. This was also caused by the disproportional sample structure: comparatively small segments were surveyed and analyzed with a sample size that was much higher than their representative market share, thus ensuring enough data for later choice estimation. Of course, these simulation results need weighting to the real segment share in a total market model. For further analysis we adapted the disproportional sample structure of study 6 (N=8,900 interviews) to the proportional market weights eliminating 3,300 interviews (N=5,600).

Introducing natural proportions into the 4 small segments, which then have insufficient sample size, led to generally worse estimates. In all 4 segments covariates were nevertheless able to improve the results. However, these were far away from the accuracy we observed with disproportional samples.



Graphic 7: Covariates in Proportional sample segments

The fourth conclusion is therefore that covariates do not allow for proportional sampling of small segment cells

TYPES OF COVARIATES AND IMPACT ON RESULTS

It is often stated that covariates work best when they add *new* information to the CBC data, and when the covariate information is strongly predictive of respondent preferences. Furthermore, it is stated that variables related to behavior and preferences will tend to be more valuable covariates than descriptive information such as demographics. Therefore we tested brand preference, past purchase, and available budget in some of our studies as potential candidates to be used as covariates. In general we could see that such added data did not result in real improvement of hit rates or MSE. If the additional information is chosen carefully it doesn't affect the HB estimation very much, but it also could degrade the results if the additional information is contrary to the original data.

Segmentation solutions based on cluster analysis of dozens of variables including preferences, attitudes, and psychographics could be valuable when introduced as categorical covariates. However, from a practitioner's point of view it is hard to know whether the estimation model will benefit. We saw in most of our studies that such complex covariates didn't improve the results compared to real market data. In nearly all of our cases, when there were changes in part-worths because of covariates, we were not able to explain the direction in which the covariates changed the results. Our attempts to use Latent Class segment membership as covariates led to small improvements of hit rates and MSE, but quite often also showed large changes in the resulting part worths. Covariates developed using only the choice data tend not to be helpful, and generally lead to over-fitting. One reason for this over fitting is that no new information from outside the CBC data is being used. The information that was already available within the CBC data was, in essence, being used twice.

In further analysis we tried to use combinations of several covariates in one model at a time. The results showed that it's generally not ideal to include several covariates without first confirming their potential usefulness by testing the distribution of the choice data compared to the additional variables. We learned that it is much better to focus on just a few covariates thereby adding relatively few columns to the covariate design matrix. One potential saving of parameters could be to treat a covariate as continuous rather than to categorize it as dummies. By using a continuous variable as a covariate one can save many parameters to be estimated without sacrificing much information. However, in many cases there is no such continuous information available. As with any multiple regression application one should carefully examine whether the covariates are influenced by multicolinearity when using more than one covariate in the HB model. But our observation was that in only one of ten studies were small improvements detected through adding combinations of covariates.

The more sparse a dataset is (e.g. relatively few choice tasks compared to the number of parameters to estimate or small samples), the more Bayesian shrinkage toward the pooled upper-level model can be expected. Therefore we examined whether covariates were most effective with sparse data sets. For those datasets with sufficient information at the individual level, the Bayesian shrinkage is already relatively small in standard HB, and covariates have a limited ability to improve the results of small sample cells. At the same time there is a risk that covariates have a negative impact on the results (i.e. MSE compared to real market data being larger). The ten studies we analyzed showed that the covariates could help to improve the results in some of our small (proportional) sample cells. However, the accuracy was still less than results from disproportional samples. Although we found that the HB shrinkage was reduced to some degree, the use of covariates did not provide the hit rates and preference share accuracy needed when communicating results to clients.

To Put it in a Nutshell

Our examination of covariates showed the following: The use of covariates is not really time saving. It neither results in shorter estimation time, nor is it fast and simple to use. The number of different estimation runs necessary to identify those covariates that improve or diminish the results required a lot of computational time and did not show any advantage against other techniques that could be used to reduce the Bayesian shrinkage.

The application of covariates reduced the precision of the estimates as often as it improved them. We found that the application of covariates is neither a fast nor an easy technique for everyday work. A lot of experience and analytical work on data structure is necessary in order to gain profound knowledge about the distribution of the data and to ensure that the applied covariates will really improve the results. The hypothesis that covariates increase the accuracy of estimations in regard to MSE, hit rates or real market data could not be confirmed in any of our ten studies analyzed.

The hypothesis that smaller samples or fewer tasks are needed when covariates are introduced could not be confirmed either. We tried systematically reducing sample sizes and numbers of tasks per respondent, in an attempt to see whether covariates helped when there was less information. But we saw that these manipulations had little effect on the quality of results, and the differences in quality occurred independently of the use of covariates. A rationale for this

phenomenon could be that in many cases we use too-long interviews and therefore get noise into our data by burdening the respondents (but this should be the topic of another paper).

Perhaps covariates could be used as a first aid kit if one observes that the data is sparse when estimating with standard HB or that there are too many parameters in the model to reach convergence in the estimates. Perhaps carefully chosen covariates adding additional exogenous information could help to improve the results in some cases. But one always should be aware that this improvement is only marginal compared to the results based on well-designed studies. Therefore we could conclude that covariates are not a "gold standard" for estimation. They could sometimes be helpful, but normally we would recommend the use of standard HB.

In most of our observations covariates in general did not improve results. In studies with large enough segment cells the covariate model converges towards same estimates as with standard HB (No influence of the covariate). Our "Gold Standard" from a practical point of view: Ensure sufficient sample size (disproportional segment cells), and use standard HB with enough iterations to assure convergence.

Covariates could improve results if we have already clearly defined clusters with groups of respondents with different multi-normal distributions on attributes in the data. Different densities in different regions of the data structure could also be a good indicator for use of covariates. We therefore suggest that one should first analyze the density structure of the common distribution of the choice data carefully and then decide whether or not to use covariates or other techniques to improve the results.

If there is a strong underlying cluster structure in the population, which was not taken into account in the sample planning and which can be identified and added as covariates to the HB estimation, this could help to improve the results. But other techniques like using proportional sampling or the weighting technique (Howell 2007) could solve the problem too. In our experience it is preferable not to add too much additional information at a time. This means trying each covariate within a single estimation before adding multiple covariates to your data.

Academics have shown that covariates can be superior in simulated environments. But in our re-analysis of ten studies, we show that using covariates can also be risky. In these studies we have had market data available to evaluate the effects of covariates, without which it would seem hard to evaluate the correct use of covariates. Covariates can in some cases improve estimates of parameters, but unfortunately not in the same amount than techniques like proportional sampling or alternative specific designs do.

REFERENCES

Fuchs, S., Schwaiger M. “Disproportionate Samples in Hierarchical Bayes CBC Analysis” in: Decker, R., Lenz, H.-J. (Eds.) „Advances in Data Analysis“ Berlin, Heidelberg (Springer), pp 441-448.

Howell, John (2007), “Respondent Weighting in HB,” Sawtooth Software Conference Proceedings, Sequim WA, pp 365-377.

Lenk, P. J., DeSarbo, W. S., Green P. E. and Young, M. R. (1996) “Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs,” Marketing Science, 15, 173-191.

Orme, B.; Howell, J.(2009) “Application of Covariates within Sawtooth Software’s CBC/HB Program: Theory and Practical Example”, Sawtooth Software Technical Papers available at www.sawtoothsoftware.com

Sawtooth Software, "More Evidence CBC Is Most Popular Conjoint Method,” Sawtooth Solutions Newsletter, Summer 2003, available at www.sawtoothsoftware.com.

Sawtooth Software (2005), “CBC/HB Technical Paper,” available at www.sawtoothsoftware.com.

Sentis, Keith and Lihua Li, “One Size Fits All or Custom Tailored: Which HB Fits Better?” Sawtooth Software Conference Proceedings, Sequim, WA, pp 167-177.

MODELING DEMAND USING SIMPLE METHODS: JOINT DISCRETE/CONTINUOUS MODELING

THOMAS C. EAGLE
EAGLE ANALYTICS OF CALIFORNIA, INC.

There are many ways to model demand, or volume. I categorize the myriad of approaches into four very general types: regression-based approaches; share, or choice, modeling approaches; joint discrete/continuous approaches; and economic models of choice. For simplicity, demand is defined here naively as the quantity of a product or products to be purchased.

Each of these approaches is reviewed conceptually at a very high level, with a discussion of their strengths and weaknesses, and discussion of a few variants that exist within each. I go into details on the joint discrete/continuous approach because that is the purpose of the paper. I do not go into detail on estimation because there are many ways to approach estimation of these models. I use four empirical data sets to compare three of the four approaches. Why three instead of all four? The reason lies in the title: "…Using Simple Methods." The economic models of choice are pretty complex and currently impossible to estimate without customized, complex, software. The other three approaches can be estimated using any current commercially available such as Sawtooth Software's CBC HB (Hierarchical Bayes) and HB Reg programs, SAS software, SPSS, Latent Gold, etc.

Any method of estimation can be used in fitting these models: Hierarchical Bayes, latent class modeling, random effects models, mixed logit, etc. *I leave the method of estimation to you, the reader.* Personally I have examined most of these estimation methods, but I traditionally use Hierarchical Bayes. Individual-level models explicitly capture more of the heterogeneity in the data, thus leading to better predictions. They have the additional benefit of user flexibility in the simulator, as new subgroups (for example male Hispanics) can be developed "on the fly" in the simulator. In aggregate level models, all of the subgroups (and this is typically a very limited number) must be defined a priori. In this paper I use Sawtooth Software's CBC HB and HB Reg programs for four of the five comparison data sets. For the last data set I use customized software because of its size.

CAVEATS

Caveat One: I am sure there are methods and approaches I have left out. I acknowledge that several of the approaches are considered 'duct tape' methods that are not theoretically elegant. My simple intent is to discuss practical solutions to estimating volumetric models that I am aware of being applied in marketing research. These models can be estimated using available software, including that from Sawtooth Software. I expect academics to be highly critical of some of the things discussed in this paper.

Caveat Two: I also apologize if I have inadvertently left out some important contributions of anyone, or if it appears I have 'borrowed' the ideas of others. It is not my intent to steal the thunder of, or contributions made by, anyone. My hope is to provoke discussion and disseminate

knowledge. I would be glad to update this document with any appropriate missing contributions of anyone.

REVIEW OF THE METHODS

Regression-based approaches:

Regression-based approaches treat volume as a continuous, or count, dependent variable whose predicted value is a function of independent variables that include product(s) attributes, respondent characteristics, or market conditions. These attributes may be brand, price, packaging, and any other attribute of the product that distinguishes it from other products. The volume may be transformed or left untransformed depending upon the nature of the volume variable and the market. These models are typically fit for each desired alternative. The simplest form of such models is:

$$Volume = f(product\ characteristics\ \&\ price)$$

Optionally the form could also include characteristics of other alternatives, the market, the observations, or combinations of all three:

$$Volume_i = f(product\ characteristics_i + price_i) + g(other\ product's\ characteristics_j) + h(market\ conditions)$$

There are a variety of such regression-based models: linear regression models, log-linear regression models, Poisson models, Negative Binomial models, Translog models, Production Function models; just to name a few. There are zero inflated variants of each of these. There are HB variants, Random-effects variants, aggregate variants, latent class variants, and even mixed variants. The use of simultaneous equations, instrumental variables, or the combination of both can be used to deal with the endogeneity issue inherent in volumetric models.

Table 1 below shows the strengths and weaknesses of this approach.

The main disadvantages of these approaches include: The components of volume, or the drivers of the change in volume (incidence, choices, and quantity) are not well differentiated; volume does not always change continuously (or change dramatically) due to changing market conditions over time or as more and more new products are introduced in the market; and adding new alternatives becomes an issue of finding the best analog from existing models in the system of volume models you built.

Some regression-based models treat each product as independent of all other products. That is, the price, or actions, of a competitor have no influence on the volume of your product. This approach is naïve because we know many products are substitutes for one another. As you change the price of one product you would expect that those products that are close substitutes (or complements) would also change. To remedy this many models incorporate ‘cross effects’ into the volumetric model of a product. For example, the volume of product A is not only a function of its own price, but also the price(s) of its competitors (or some weighted market price index). Issues with these approaches include: you have to limit the number of ‘cross effects’ in large markets; the ‘cross effects’ are very unstable and can give you incorrect signs because of multicollinearity among the ‘cross effects’; there could be market price effects not captured correctly using price ‘cross effects’; and they cannot be easily adapted to the addition or

subtraction of new products. Any of you who have ever fitted the 'Mother Logit' MNL model know how difficult fitting models with 'cross effects' can be.

Table 1: Strengths and Weaknesses of the Regression-based Approaches

Strengths	Weaknesses
• Easy to fit • Variety of model forms • Substitution not subject to MNL model's independence of irrelevant alternatives property • Volume not constrained to maximum • Budget constraints can be implemented • Best choice for aggregate data where data generating process is unclear	• Components of volume (incidence, choice, quantity) not well differentiated • Extreme values have strong effects • Without 'cross-effects' substitution is non-existent • Model fits are inconsistent • Typically, no 'satiation' effects or diminishing returns captured • Error terms across alternatives are correlated

Typically the fit of these models is generally quite poor (especially aggregate forms of such models). Aggregate model R-square values in the .20 to .30 range, or lower, are not uncommon. HB regression, or random effect, models improve the fit, but they are generally only improved by having random effect intercept terms. Other terms such as the impact of price, or price cross effects, can be messy to say the least.

As a result of these drawbacks many practitioners have moved to using market share, or choice models, to capture the patterns of substitution among products and modified these to model volume.

Choice modeling approaches to modeling volume

The traditional choice model is designed to model the patterns of substitution among products. Depending upon the choice model being used the choice model can also handle the addition and deletion of products in the market easily (i.e., the MNL choice model). The primary drawback of choice models is they predict probabilities of choice; not volume. As such, as you sum the predicted probabilities of the products in the market, the probabilities sum to 1.0 -- always. It is a 1.0 sum game. But volume is continuous. So how can one model volume using a choice model?

One approach is to rebase the volume estimates provided by respondents to proportions. One simply divides the volume assigned to a single alternative by the total volume assigned across all alternatives in a choice task. This may also be rebased to a constant sum. A problem arises if volume changes from task to task. For example, there may be some tasks where zero volume is assigned; other where the volume may vary widely. Choice modeling approaches to modeling volume must account for the variability in total volume across tasks.

This is accomplished by adding another alternative to each choice task called a "Synthetic None." (I borrow this term from David Lyon.) That is, there is an estimatable probability that some respondents will not buy a constant number of units. For example, as the prices of all

products rise in a market one might expect total volume across alternatives to drop. This drop in volume is accounted for in a choice-base volume model by increasing the volume, or share of volume, assigned to the "Synthetic None" alternative. That is, the probability of the "Synthetic None" alternative would grow as all products raise their prices. How do we assign volume, share, or a probability to this "Synthetic None" alternative? Typically we find a maximum expected volume which can be used to assign volume to the "Synthetic None" alternative.

Prior to estimation of the choice/share volumetric model, the researcher totals the alternative specific volumes in each task. The researcher scans across all the tasks each respondent completed. The maximum total volume found across these tasks is designated as a benchmark volume, or Maximum Expected Volume (MEV). This is unique to each respondent, so the volumes are individual-level expectations. There may also be many other variants of estimating this maximum expected volume. Using the MEV we can now transform each volume for each alternative for each choice task into something we can use to fit a choice model. We can also now assign volume to the "Synthetic None" alternative.

We use the MEV as a fixed total volume for each choice task a respondent completes. We sum the volumes assigned across all alternatives in a task. If this sum is less than the MEV we assign the difference to the "Synthetic None" alternative. If this sum equals the MEV, the volume assigned to the "Synthetic None" alternative is zero. Now every choice task has a constant sum so we can use those, or rebased them to proportions, in a traditional choice model. If we are using the CBC HB MNL program, this model is a MNL model.

The volumetric choice/share model is estimated. A prediction simulator is built. Predicted probabilities are calculated in the simulator and these are multiplied by the MEV to translate the probabilities back into units of volume. Because the "Synthetic None" alternative has a predicted probability associated with it, the actual volume going to the other products will be less than the MEV. As the other products raise (or lower) their prices, the volume associated with them will drop (or grow). This gives us a "share" model of demand.

Basically what we have is a model:

$$Volume_i = Prob_i * MEV$$

Where:

$$Prob_i = f(brand_i, features_i, price_i)$$

$f()$ is any functional choice model form.

Technically this could be an MNL model, a nested logit model, a MNL probit model, any form of a choice model that predicts a probability for each alternative in a choice set.

Table 2 gives some of the advantages and disadvantages to this approach:

Table 2: Strengths and Weaknesses of the Choice Modeling Approaches

Strengths	Weaknesses
• Easy to fit • IIA substitution • Non-independent estimates of volume • Stable models • No extreme predictions • Model fit from good to very good	• Volume may change non-continuously • Volume is capped at MEV • "None" treated as IIA with other alternatives • Extreme outliers can affect estimates • Not handling true 'multiple discreteness' • No 'satiation'

These are easy to fit because traditional Multinomial Logit (MNL) choice models can be used to fit them; they capture the substitution effects especially well (as choice model are designed to do); and they are fairly accurate. They are excellent models when demand is not highly variable across choice tasks. When running a naïve R-square of the predicted volume on actual volumes we can achieve higher R-squares than the naïve regression approaches. I have seen R-square values in the neighborhood of .6 and higher.

There are several drawbacks and some can be major. If volume is highly variable across tasks for a respondent, across respondents, or dramatically changing in the market then the estimation of the actual volume going to all non-"Synthetic None" alternatives can be off. That is, this approach works best when volume change is limited. Unless the choice model of volume allows the "Synthetic None" alternative to reach a zero probability, the MEV will never be achieved. Another drawback is that the pattern of substitution between the "None" alternative and all other alternatives is seriously miss-specified if simple MNL models are used. More advanced choice models, such as the nested logit, or GEV model, should be used to capture the substitution patterns among the "Synthetic None" alternative and other alternatives. Most practical research does not deal with the either of these issues. Nevertheless, these models always result in better predictions than using the naïve MNL choice model. Personally I have fit the nested logit version of this model using a multi-step HB approach. For the purposes of the case studies presented later I use the MNL model.

Joint Discrete/Continuous volume models

This approach combines the advantages of the choice/share model with the advantages of the regression-based volume models. Much of what I describe below is found in chapter 5 of Ken Train's book, Qualitative Choice Analysis: Theory Econometrics, and an Application to Automobile Demand (1986) and in the Hausman, et al, (1995) paper in the Journal of Public Economics. The approach described below uses a sequential estimation of the model. Several people are currently working on building full information, simultaneous estimation, versions of these models.

There are two stages to fitting a joint discrete/continuous volumetric model: the fitting of an allocation, or share, choice model; and the fitting of either a total task volume model or alternative specific regression-based volumetric models. I examine each in turn.

Step 1: Fit an allocation choice model

In these models, an allocation choice model(s) is first fit with, or without, the inclusion of a "None" alternative (if the 'None' alternative is included I fix its utility to zero). Notice I am not using the term "Synthetic None" in this case. The only time a "None" alternative is added to the model is when the total volume assigned across all alternatives is zero. If all choice tasks have non-zero volumes then the inclusion of a "None" alternative is unnecessary. If some tasks have zero volume assigned to alternatives then we must add a "None" alternative. The "None" alternative would have a zero volume/allocation assigned to it in every choice task with a non-zero volume assigned to at least one alternative. When a choice task has a zero volume then we assign a single unit to the "None" alternative. The volume for each alternative is then rebased such that they sum to 1.0. Using these data we fit the desired choice model. This share, or allocation, choice model captures the substitution among products and the effects of the entry/exit into the market of products (if desired). Notice that the only difference between this stage and the choice modeling approach to modeling volume is the handling of the "None" alternative. Typically I would fit a nested logit model to these data precisely because I do not expect the substitution between the "real" alternatives and the "none" to be IIA. For the purposes of the case studies presented later I use the MNL model.

$$Prob_i = f(brand_i, features_i, price_i)$$

Where:

$f(brand_i, features_i, price_i)$ is any form of a share, or allocation, type of choice model.

There are two nice outcomes of the allocation choice model: the predicted probability of each alternative and the denominator of the choice model. The denominator of the MNL model is the sum of the exponentiated utilities across all alternatives in the task. I call this the expected net utility of the all alternatives in the task. In the choice modeling literature this denominator is also called the inclusive value. For the purposes of the case studies presented I use a MNL model.

For the next stage of estimation we use either the predicted probabilities for the alternatives in the fitting of the alternative specific regression-based volumetric models, or the expected net utility term (after taking the natural logarithm of its value) in the regression-based total task volumetric model. These are the two approaches I have used and describe below.

Step2a: Fit alternative specific volumetric model(s):

We fit a regression-based model (any of the common variants is OK to use: e.g., linear regression, count, log-transformed continuous) to each alternative using the predicted probability in place of the intercept term. One can optionally add an intercept term, but I have found the value of this to be very unstable and usually near zero. One could also add a price index to each alternative's model to account for market conditions not captured by the predicted probabilities. Other terms may be added as well (e.g., promotions, etc.). Train (1986, pp. 91-97) discusses some adjustments that can be made to the price index parameter to account for bias that occurs, but these are usually minor adjustments.

The form of the alternative specific model is:

$$Volume_i = \beta_i \widehat{Pr_i} + \gamma_i PriceIndex + \sum_{k=1}^{m} \delta_k X_k$$

Where:

$\widehat{Pr_i}$ is the predicted probability of alternative i from stage 1
$PriceIndex$ is an optional measure of market price (e.g., share weighted price index)
X_k are k to m other effects that are not captured in the predicted probability for alternative i
$\beta, \gamma, and\ \delta$ are parameters to be estimated; β constrained to be positive; γ constrained to be negative.

Note there is no intercept term. Train (1986) replaces the standard intercept with the $\beta(Pr_i)$ term (though I have sometimes included the intercept). I should note that Train (1986) has a convincing argument that the price index parameters need to be corrected using an adjustment factor if a price index is included.

Note that 'cross effect' terms (e.g., the price of alternative j on the volume of alternative i) needed in the regression-based volumetric models to capture substitution are not required! The predicted probability of the alternative of interest ($\widehat{Pr_i}$) i changes when you change the attributes of other alternatives (alternative j) through the MNL model. Thus, a form of 'cross effect' is already present in the use of the predicted probability. A pretty stable model… unless you have a lot of alternatives….

I have not only fit these models separately for each alternative, but also by combining the separate alternative-specific regression-based volumetric models into a single large multiple regression by appropriately diagonalizing the X matrix (independent variables) and stacking each alternative's model data into a single data set.

Step2b: Fit an overall task volumetric model:

An alternative approach is to fit the sum of the volumes across all products, or total task volume, using the natural logarithm of the net expected utility (the MNL denominator or the inclusive value [IV]) from the choice/share models fit in stage 1 as the independent driver(s) of volume. The form of this model is quite simple:

$$Total\ Task\ Volume_k = \alpha + \beta \ln(IV_k)$$

Where:

$\ln(IV_k)$ is the natural logarithm of the predicted MNL denominator (the IV, inclusive value) for task k
$a\ and\ \beta$ are parameters to be estimated and β constrained to be positive.

As the expected net utility (IV) across all products grows then volume should grow. Competitive impacts on a product are captured in the share choice model (stage 1), which results in a potential change in the expected net utility of a product, which may lead to a change in predicted total task volume. Once a total task volume is predicted this is multiplied by the alternative specific predicted probabilities (excluding the "None") to distribute the total task volume across alternatives.

As with the earlier approaches Table 3 shows there are strength and weaknesses.

Table 3: Strengths and Weaknesses of the Joint Discrete/Continuous Approaches

Strengths	Weaknesses
• Predicted volume not capped • Very stable model • Any type of choice model • Any type of volume model • Potential misspecification of "None" avoided • Total volume can be modeled with some satiation: log(IV) • Can be used with aggregate, disaggregate revealed preference or stated preference data	• Potential unspecified bias • Sequential estimation inefficient • Increased complexity – more models • No direct 'satiation' effects unless built into volume model • Extreme outliers can affect volume estimates • Not handling true 'multiple discreteness'

The advantages of this approach over the choice modeling approach include: volume can grow rapidly or slowly; it can grow with diminishing returns (grow at a decreasing rate) as net expected utility rises; substitution effects are captured via the share choice model (stage 1); there may not be a "None" alternative in the share choice model (hence no "Synthetic None" MNL model miss-specification); volume predictions are not constrained by the use of a maximum expected volume; and they are easy to estimate using existing software. We have a great deal of flexibility in fitting the volumetric stage of the approach.

The drawbacks include: the sequential estimation of staged models; the possibility that, as the number of products increase over what was tested, the prediction of volume can still grow faster than what might be expected; the potential for some unspecified bias in the estimates (similar to that described by Train – I'll leave that to more experienced statisticians than me to consider); more complex simulator development; and the increased complexity of the modeling – two stages rather than one.

Specification of the regression-based model is critical to getting good fits. Personally I have seen these models produce naïve R-square values exceeding the other approaches when volume is highly variable or discontinuous across tasks. When I have compared this approach to the choice modeling approach described above it always does as well as, if not better than, the choice modeling approach even when volume is invariant across tasks (in these cases the parameters for the $\ln(IV_k)$ are near zero and the intercept terms are significant).

Economic models of choice

This is a catch-all category consisting of much of the work of Bayesian modelers in marketing research. The types of models to which I am referring are those that maximize a direct utility function subject to budget and other constraints and that use volume as part of the model's formulation. The Allenby references (Satomura, Kim, and Allenby, 2010; Neeraj, Allenby, and Ginter, 1998; and Kim, Allenby, and Rossi, 2007) are examples of these kinds of modeling. These models are theoretically very elegant and are designed to address some of the weaknesses of the volume models described above: for example non-linearity of the utility function,

integration, heterogeneity, multiple discreteness, the simultaneous brand/quantity decision, complementary good, and satiation. However, these models are very complex to fit and software to fit them is not readily available for the common practitioner to use. They are not suited for aggregated data (such as Nielsen data) and they often have serious convergence issues when the problems become complex. In some cases the models use volume to predict the probability of choice (share), but not volume itself.

An example of these models is the Satiation model (Satomura et al, 2010). Utility is modeled as a nonlinear function that allows for satiation; that is, diminishing marginal utility.

$$u_j(x_j) = \beta_j(x_j + 1)^{\delta_j}$$

Where:

u_j is the marginal utility of alternative j
x_j is the quantity (volume) of alternative j demanded
β_j is the baseline level of marginal utility associated with δ_j
δ_j is the satiation associated with alternative j

These models are estimated using Hierarchical Bayesian methods. They involve the estimation of a direct utility function, as opposed to the indirect utility function found in most MNL models. They are customized to address very specific issues in marketing and in the modeling of consumer behaviors. And they sometimes deal with multiple discreteness and handle multiple constraints on behavior. Like all the approaches I have described above they do have their strengths and weaknesses.

The advantages of these approaches are their theoretical elegance (see Table 4 below). Until the software becomes readily available we cannot examine their ability to work in a variety of settings. More often than not, they are tested with Monte Carlo simulations and on simple, very limited, data sets. Where variants of them have been tested in complex problems many practitioners face, we have found they do not converge easily, they require enormous amounts of computing time, and simulator based predictions of scenarios require lengthy Monte Carlo simulations. In some cases adding alternatives to a scenario requires making assumptions about the added alternative that can be challenged.

Table 4: Strengths and Weaknesses of the Economic Models of Choice Approaches

Strengths	**Weaknesses**
• Flexible forms for direct utility functional forms • Includes income and budget effects • Integrated treatment of "no choice" or "outside good" • Parsimonious specification facilitates interpretation of heterogeneity	• Software not readily available ◦ Especially for industrial size problems • Long burn-in times for MCMC • Convergence problems • Even for simple models • Not good for aggregate data • Complex and lengthy simulation required

Because the software is not readily available, the case studies presented in the next section do not include any of these economic models of choice.

COMPARISON OF APPROACHES

In this section of the paper we examine the predictive capabilities of the regression, choice model, and total volume version joint discrete/continuous (step2b) approaches on a collection of 4 simple data sets. After these four data sets are compared we examine the choice modeling and joint discrete/continuous approach on a more complex data set; dropping the regression approach because of the data set's complexity makes regression based models impractical.

Four small volumetric data sets were provided to this author by Sawtooth Software. These data sets were provided to Sawtooth Software by Hernán Talledo, director general of Grupo Epistéme. We are grateful for his permission to use these data. The attributes are disguised (unlabeled). Each data set is related to health care in the pharmaceutical arena. Three alternatives/products/treatments made up each task. The respondents were asked to indicate the number of patients who would receive the product/treatment. The data sets differed in terms of the number of attributes, their levels, the number of tasks each respondent completed, and the number of respondents. The designs were built using the randomized assignment of attribute levels, so a classic fixed holdout task is not available.

I used two metrics for comparing the predictive results across these four data sets: mean absolute error (i.e., abs[actual volume – predicted volume]) and R-square. The actual volume was compared to the final predicted volume for both metrics. Because the designs were random assignments of levels to each respondent, a common holdout task was not available for us to compare aggregate level market share or volume predictions. This means the holdout task is different for every respondent. As a result these metrics are computed at the individual respondent task level. The metrics are the individual respondent task-specific alternative-specific actual volumes compared to the predicted volumes for the same. This is a much more stringent test of model accuracy than those based upon aggregate data. For each of the four data sets a single task (e.g., the respondents 7^{th} task) was randomly selected for comparison purposes. This holdout task was varied across the four data sets.

For each of these four data sets the estimation parameters were held constant:

- 10,000 burn-in iterations were used for all HB estimations (the regression models did not require such a long burn-in, but we did so to maintain consistency)
- 10,000 sample iterations
- Mean posterior parameter estimates were used in all predictions
- All attributes were treated as part-worth effects with no constraints
- No upper level model covariates were used

The design specifications for each of the four data sets are in table 5 below:

Table 5: Design Specifications for the 4 Data Sets

Study designation	A	B	C	D
Number of attributes	9	7	7	8
Design characteristics	$6^2 \times 5 \times 4 \times 3 \times 2^4$	$6 \times 5 \times 4^3 \times 3^2$	$7 \times 6 \times 4 \times 3^2 \times 2^2$	$7 \times 4 \times 3 \times 2^5$
Is price an attribute?	Yes	Yes	No	No
Prohibitions in design?	Yes	Yes	Yes	No
Task per respondent	12	15	12	12
Number of respondents	134	302	133	253

Table 6 shows the predictive fit measures for each of the four data sets based upon the data upon which the models were estimated. Joint D/C refers to the joint discrete/continuous volumetric approach. Note we show the fit measures for the choice modeling components of the choice modeling approach and the joint D/C approach for your information.

Table 6: Model Fits on Estimation Data for the 4 Data Sets

		Model Fit on Estimation Data							
		R-Square				MAE			
Approach	Model	A	B	C	D	A	B	C	D
Regression	Alt 1	0.909	0.877	0.912	0.862	4.038	5.821	3.740	7.052
	Alt 2	0.912	0.879	0.850	0.876	3.885	5.968	4.012	6.519
	Alt 3	0.891	0.876	0.876	0.836	4.007	8.758	3.454	7.306
Choice model	Choice*	0.767	0.766	0.861	0.752	0.066	0.056	0.061	0.061
	Final Volume*	0.838	**0.842**	**0.890**	**0.868**	3.950	**5.788**	**2.957**	**6.177**
Joint D/C	Choice	0.906	0.861	0.914	0.870	0.058	0.070	0.059	0.066
	Total Volume	0.916	0.766	0.879	0.847	7.206	19.856	7.35	17.732
	Final Volume	**0.882**	0.712	0.825	0.815	**3.656**	7.641	3.632	7.175

* = using only alts 1-3

The R-square values are on the left side of the table and the MAEs on the right. At first glance you would think the regression models are the best fitting models based upon the R-square values and some of the MAEs. However, you should note that these are individual-level alternative specific models designed to maximize the R-square which minimizes the MAE as well. These models, however, were fit without cross effects of other alternatives attributes on each alternative. As such they have no substitution effects. Any change in one alternative's attributes has no impact on the other alternatives volume. We did not put any cross effects into these models because we did not know what the attributes were and price was not always available. Moreover, the mean across these regression models is approximately the same as the choice model and joint D/C approaches.

The red figures highlight the best fits among the choice modeling and joint D/C approaches. In three out of the four data sets the best fit is from the choice modeling approach. This might be expected simply because the joint D/C approach is attempting to fit two interdependent models (two sources of error). Nevertheless, BOTH approaches fit the data remarkably well. The magnitude of differences is relatively small. The next table (Table 7) shows similar results for the holdout tasks. Figures 5 and 6 in the appendix show the scatter diagrams of actual volume against predicted volumes for data set A. The other data sets have diagrams that are very similar.

Table 7: Model Fits on Holdout Data for the 4 Data Sets

		Model Fit on Holdout Task							
		R-Square				MAE			
Approach	Model	A	B	C	D	A	B	C	D
Regression	Alt 1	0.666	0.800	0.497	0.704	9.433	8.776	7.360	10.932
	Alt 2	0.471	0.676	0.612	0.675	7.114	11.064	6.076	9.32
	Alt 3	0.537	0.525	0.498	0.542	7.057	10.018	6.199	10.337
Choice model	Choice*	0.896	0.292	0.280	0.491	0.053	0.101	0.145	0.082
	Final Volume*	**0.917**	**0.677**	**0.419**	**0.687**	**3.215**	**10.477**	**7.006**	**8.577**
Joint D/C	Choice	0.916	0.147	0.338	0.429	0.047	0.164	0.193	0.144
	Total Volume	0.931	0.870	0.844	0.892	7.695	24.849	8.466	16.115
	Final Volume	0.887	0.587	0.394	0.665	3.664	12.14	7.632	9.449

* = using only alts 1-3

The choice modeling approach wins in all four cases. In three out of four cases the holdout MAEs are larger than those found in the estimation data.

If we were to stop here one might conclude that the choice modeling approach is better than the joint D/C approach. It is easier to fit and performs slightly better. That might be a hasty conclusion. While it was not the intention of the original research, the design of these tasks was especially suited for the fitting of the choice modeling approach. The tasks were treatments to which respondents (i.e., physicians) we asked to assign an open-ended number of patients. The numbers of patients a physician sees with the conditions for which these treatments are capable of treating are fixed for each physician at the time of task. As such, it is quite possible physicians were allocating their fixed number of patients such that those who were not given one of the three treatments were being mentally assigned to an "All Other Treatment" category. The fact that the tasks were all small in terms of numbers of attributes and alternatives, fixed in terms of available treatments, and relative generic (there were no alternative specific attributes) makes this task highly amenable to being fit well with the MNL specification of the choice modeling approach. Moreover, these same conditions mask the capabilities inherent in the joint D/C approach.

A More Complex Comparison

For this reason I decided up pull a more complex data set from one of my recent projects where I used the joint D/C approach. It was a pricing and new product introduction study done for a CPG firm. For obvious reasons I cannot divulge the exact nature of the products, the client, nor the specific attributes and results. The objective of the study was to find the optimal portfolio of existing and new products with appropriate pricing in the face of existing competitive

products. 38 SKUs were manipulated using a presence/absence design. Some SKUs exist in the marketplace and always present (5 for the client and 17 for the competition). The remaining new SKUs were all from the client. The final design showed a range of 23 to 38 alternatives in each task. Price was manipulated for each SKU. Some prices were linked across SKUs so that illogical combinations of prices would not appear. The data set had 1000+ respondents who completed 11 tasks including a holdout task.

For the comparison I dropped the regression approaches because I felt attempting to fit models with up to 37 'cross effect' price parameters would be futile.

The final design consisted of 88 tasks blocked into 8 blocks of 11 tasks each. The tasks were randomized within and across respondents. One holdout task was pulled from each respondent prior to estimation. This holdout task differed across blocks.

I used customized software to fit these models because the Sawtooth Software product could not estimate the model in a reasonable amount of time. The estimation times increased because of the inefficiency of the Sawtooth Software CBC HB in fitting models with any large number of upper level model covariates (this is being looked into as this paper is being written). In addition, the Sawtooth Software HB Reg program does not handle upper level model covariates. The two approaches were run using similar estimation parameters:

- There were 16 upper level model covariates
- 60 lower level model parameters (the utility functions of the MNL models)
- 20,000 burn-in iterations
- 10,000 posterior draws saving every 10th iteration
- Posterior mean parameters were used for predictions

Table 8 below shows the comparison of results for the choice modeling and joint D/C approaches. In this particular example I also fit the alternative specific option (step 2a) because that was the final model used to provide results to the client.

Table 8: Model Fits on Estimation and Holdout Data for the Complex Data

		Fit Measures					
		Estimation			Holdout		
Approach		R-Square	MAE	VolShMAE	R-Square	MAE	VolShMAE
Choice model	Choice*	0.514	0.022		0.466	0.021	
	Final Volume*	0.376	0.207	0.0103	0.316	0.197	0.0087
Joint D/C Total Volume	Choice	0.541	0.029		0.408	0.029	
	Total Volume	0.782	1.693		0.750	1.820	
	Final Volume	**0.433**	**0.176**	**0.0071**	**0.367**	**0.173**	**0.0061**
Joint D/C Alt. Specific Volume	Final Volume	*0.592*	*0.127*	*0.0041*	*0.422*	*0.131*	*0.0042*

There are three fit statistics presented: the R-square and MAE measured exactly the same as in the previous comparison, and an aggregate volumetric market share MAE (VolShMAE). The VolShMAE consists of aggregating the individual respondent volume numbers into aggregate total sample volumes for each alternative and converting them into volumetric market shares by dividing the alternative specific aggregated volume by the total aggregated volumes across all alternatives. The VolShMAE is then computed as abs[actual volume market share – predicted volume market share]. The left side of the table shows the results for the estimation data and those on the right are the results for the holdout tasks.

In this data set the total volume joint D/C approach clearly outperforms the choice modeling approach on all three metrics. The relative improvement in R-square is 15% (=.433/.376). For MAE, the relative improvement is 15% (=.176/.207). The improvement in volumetric market share MAE (VolShMAE) is 31% (=.0071/.0103). The holdout tasks all show substantial decreases in the R-square metrics, but not the MAEs. If you examine the choice modeling component metrics for the two approaches you will see they are also very similar. It is in the estimation of volume that that joint D/C approach outperforms the choice modeling approach. Clearly the use of the "Synthetic None" is inappropriate for these volume data.

The magnitude of the volumetric market share MAEs should not alarm anyone because the number of alternatives in a task directly affects any MAE measure. The more alternatives you have the higher the number of alternatives with zero volumes, hence the lower the MAEs. In volumetric comparisons such as these the absolute MAE is not the meaningful metric. It is the relative difference in MAEs.

The interesting conclusion is the marked improvement of the alternative specific joint D/C model over both the total volume joint D/C and the choice modeling approach. In terms of the three metrics the improvements are:

- Alt. Spec. joint D/C volume over choice modeling R-square: 57% (=.592/.376)
- Alt. Spec. joint D/C volume over total volume joint D/C R-square: 37% (=.592/.433)
- Alt. Spec. joint D/C volume over choice modeling MAE: 39% (=1 - .127/.207)
- Alt. Spec. joint D/C volume over total volume joint D/C MAE: 38% (=1 - .127/.176)

The alternative specific joint D/C approach is the clear winner in this comparison. I believe the difference in results (from the initial 4 data sets and this data set) is a function of the task complexity. This data set is a much more complex task. There are many more alternatives, many with zero volume. Prices are being manipulated as well as the presence and absence of alternatives. And, most importantly, the task is not an obvious allocation task as were the previous comparison's data sets.

SUMMARY

Joint discrete/continuous volumetric modeling is a valid and flexible approach to modeling complex volume model. We summarize our conclusions in Table 9 below:

Table 9: Summary Comparison of the Choice Modeling and Joint Discrete/Continuous Approaches

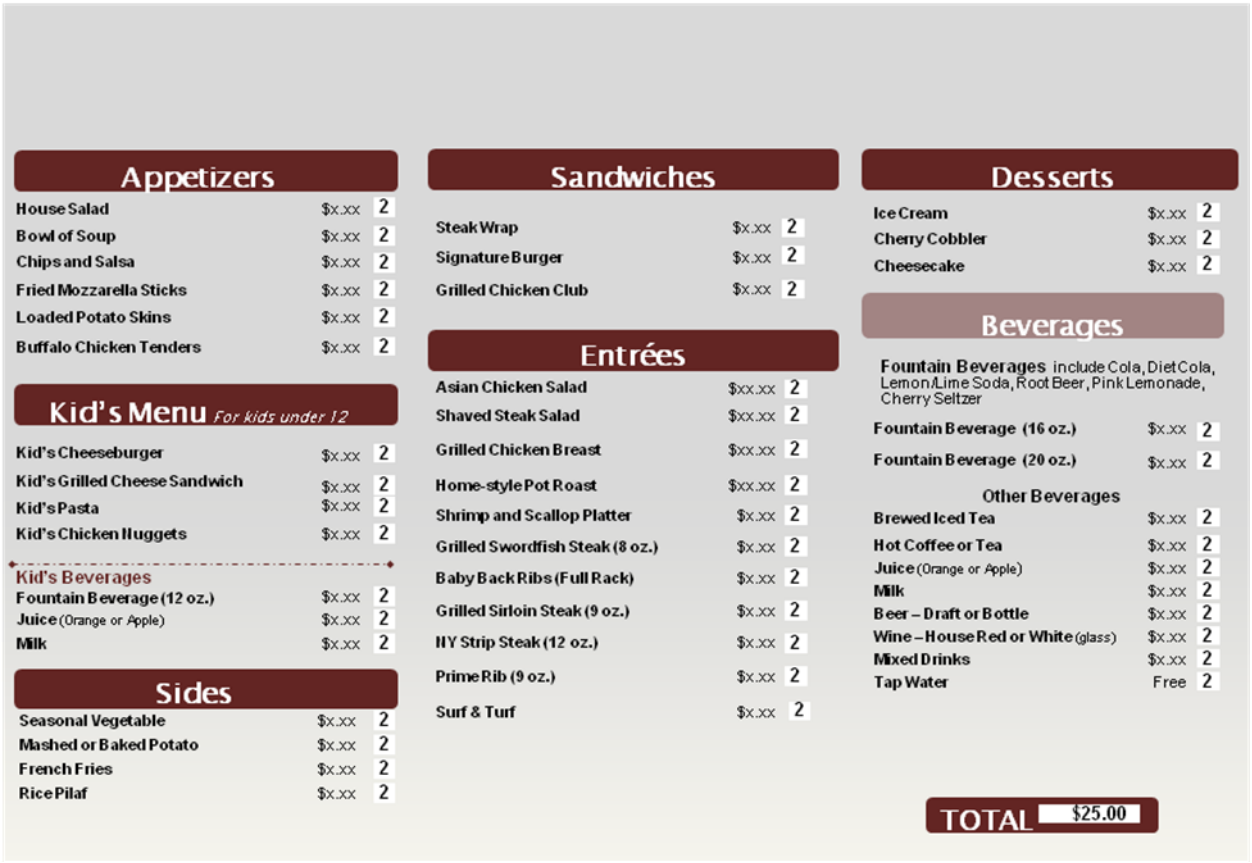
Choice Modeling Approach	Joint Discrete/Continuous Approach

• Easy to fit – one model fits all…
• Volume constrained to MEV
• Stable models
• There is no real volume model
• In MNL substitution is IIA with the "None" AND all other alts
• For simple models and allocation-like data
  • Fits well, if not better than joint D/C approach
  • Application more limited to smaller problems
	• Easy to fit – 2 stages
	• Volume unconstrained
	• Stable models
	• Any volume model with expanded modeling capabilities can be used
	• Substitution is IIA only among alts in task – if MNL model used
	• For complex models and data
	  • Fits better than choice modeling approach
	  • Offers more flexibility in modeling

More complex modeling adventures

To quickly demonstrate the flexibility of the joint discrete/continuous approach we pose to you the modeling of a full service restaurant menu. An example of such a menu is presented in Figure 1 below.

Figure 1: An Example Full-service Restaurant Menu



We have complementary and substitutable items on the menus. We have an implied budget constraint. We have potential satiation. Using a choice modeling volumetric approach would be very difficult in this situation. We might explode the menu into all possible combinations of menu items and estimate the choice modeling volumetric approach model. But this would lead to thousand of combinations to model. Or we might break the menu up into categories and fit separate choice modeling approach volume models to each category. But this would require a large number of 'cross effects' within each category's choice modeling volumetric model which can be problematic. It would also involve the use of the "Synthetic None" alternative.

Using a joint discrete/continuous approach can simplify this problem. We take advantage of the categorization idea mentioned immediately above to break the menu into its logical food item categories: appetizers, entrées, desserts, etc. We fit each category as a separate joint discrete/continuous volumetric model using the steps we outlined above for the total volume approach (step 2b). Thus, we would fit the MNL, or nested logit, model as we described in stage 1 of the joint D/C approach for each category. We compute the total expected utility (inclusive value; i.e., the IV) for each menu category. However, when fitting of the total volume regression step (2b) for any specific category using these total expected utilities ($\ln(IV_k)$) we add the other menu category's total expected utilities into the model as cross effects. The signs of these cross effects could be negative, which suggests substitutability across categories, or positive, which suggests complementarity across menu categories. The results volumetric portions of the model would look like:

$$Volume_k = f\left(\beta_k \ln[IV_k] + \sum_{m=1}^{M} \beta_m \ln[IV_m]\right)$$

Where:

$Volume_k$ is the volume assigned all products in menu category k,

$\ln[IV_k]$ is the natural logarithm of the total expected utility (predicted denominator of the MNL model) for category k,

$\ln[IV_m]$ is the natural logarithm of the total expected utility (predicted denominator of the MNL model) for category m, and

β_k and β_m are parameters to be estimated

This model will have more stable cross effects than those fit using the choice modeling approach to the menu. In addition, it gives us the flexibility to fit zero inflated models which would be logical given the number of zero volumes that will appear on the menu. We can also incorporate budget constraints into the volume portion of the joint D/C approach which we could not do in the choice modeling approach. I have fit several of these menu board models using the joint D/C approach with great success every time I have compared them to choice modeling approaches.

REFERENCES

Classic sources on Joint Discrete/Continuous Modeling

Hausman, J.A., G.K. Leonard, and D. McFadden (1995), "A Utility-Consistent, Combined Discrete Choice and Count Model Assessing Recreational Use Losses Due to Natural Resource Damage," Journal of Public Economics, 56, 1-30

Train, Ken (1986), Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand, Cambridge, MA: The MIT Press. A pdf file for the entire book may be found at: http://elsa.berkeley.edu/~train/qca.html

Hanemann, W. M. (1984), "Discrete/Continuous Models of Consumer Demand," Econometrica, 52, 541-61

Multiple constraint choice model paper

Satomura, Takuya, Jaehwan Kim, and Greg Allenby (2010), "Multiple Constraint Choice Models with Corner and Interior Solutions," working paper.

Allenby and Garratt's trade-up model

Allenby, Greg, Mark J. Garratt and Peter E. Rossi (2010), "A Model for Trade-Up and Change in Considered Brands," Marketing Science, 29 Jan-Feb, 40-56

Allenby and Rossi's homothetic and non-homothetic volume models

Rossi, Peter, Greg Allenby and Robert McCulloch (2005), Bayesian Statistics and Marketing, West Sussex, England, John Wiley & Sons, Ltd.

Other HB volumetric models (for example)

Neeraj, Allenby, & Ginter (1998), "A Hierarchical Bayes Model of Primary and Secondary Demand," Marketing Science, 17, 29-44

Chintagunta, Pradeep K. (1993), "Investigating Purchase Incidence, Brand Choice and Purchase Quantity Decisions of Households," Marketing Science, 12 (Spring), 184-208

Kim, Jaehwan, Greg M. Allenby and Peter E. Rossi (2007), "Product Attributes and Models of Multiple Discreteness," Journal of Econometrics, 138, 208-230

APPENDIX

Figure 2: Data set A's distribution of alternative specific volume

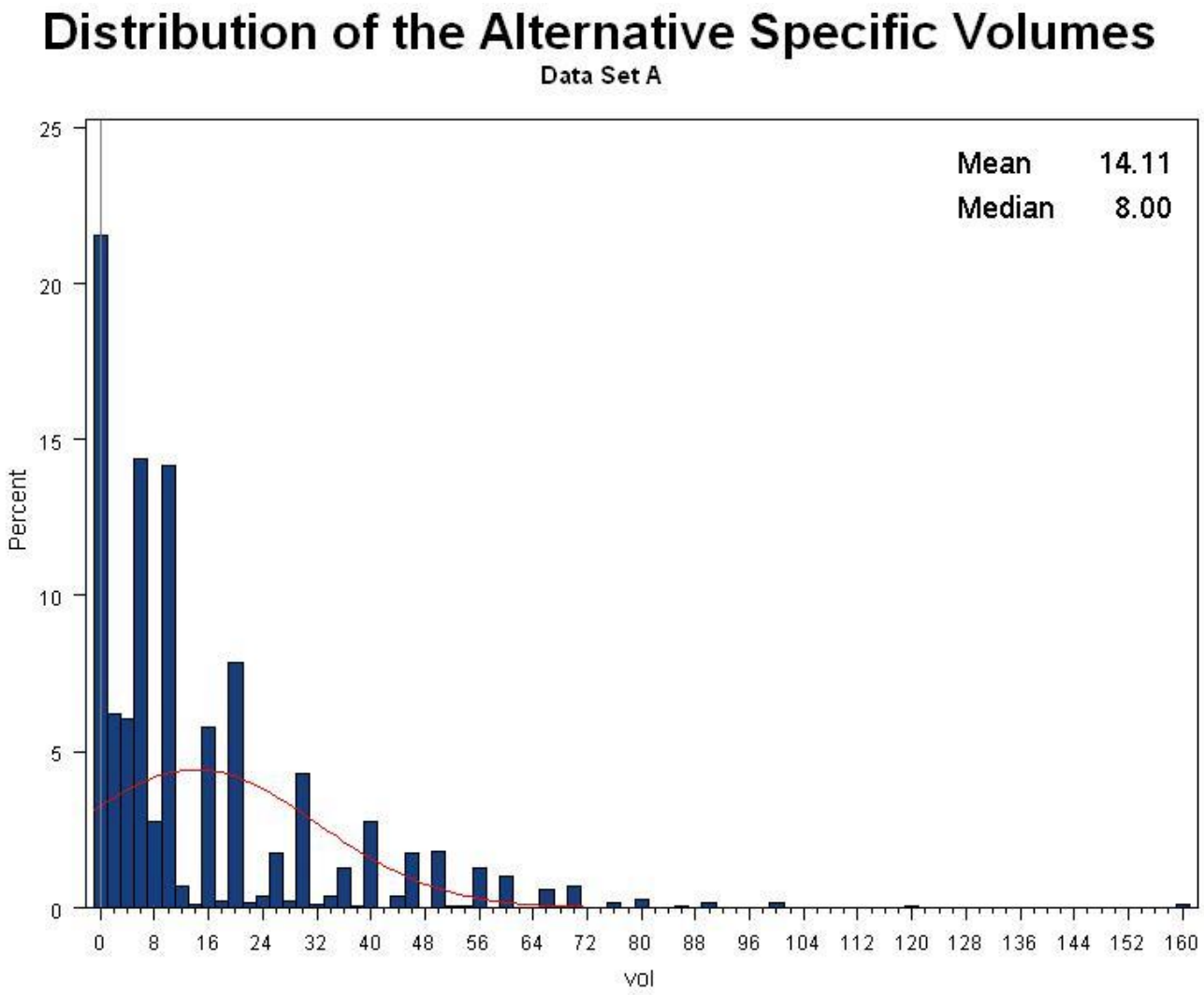


Figure 3: Data set A's distribution of total task volumes

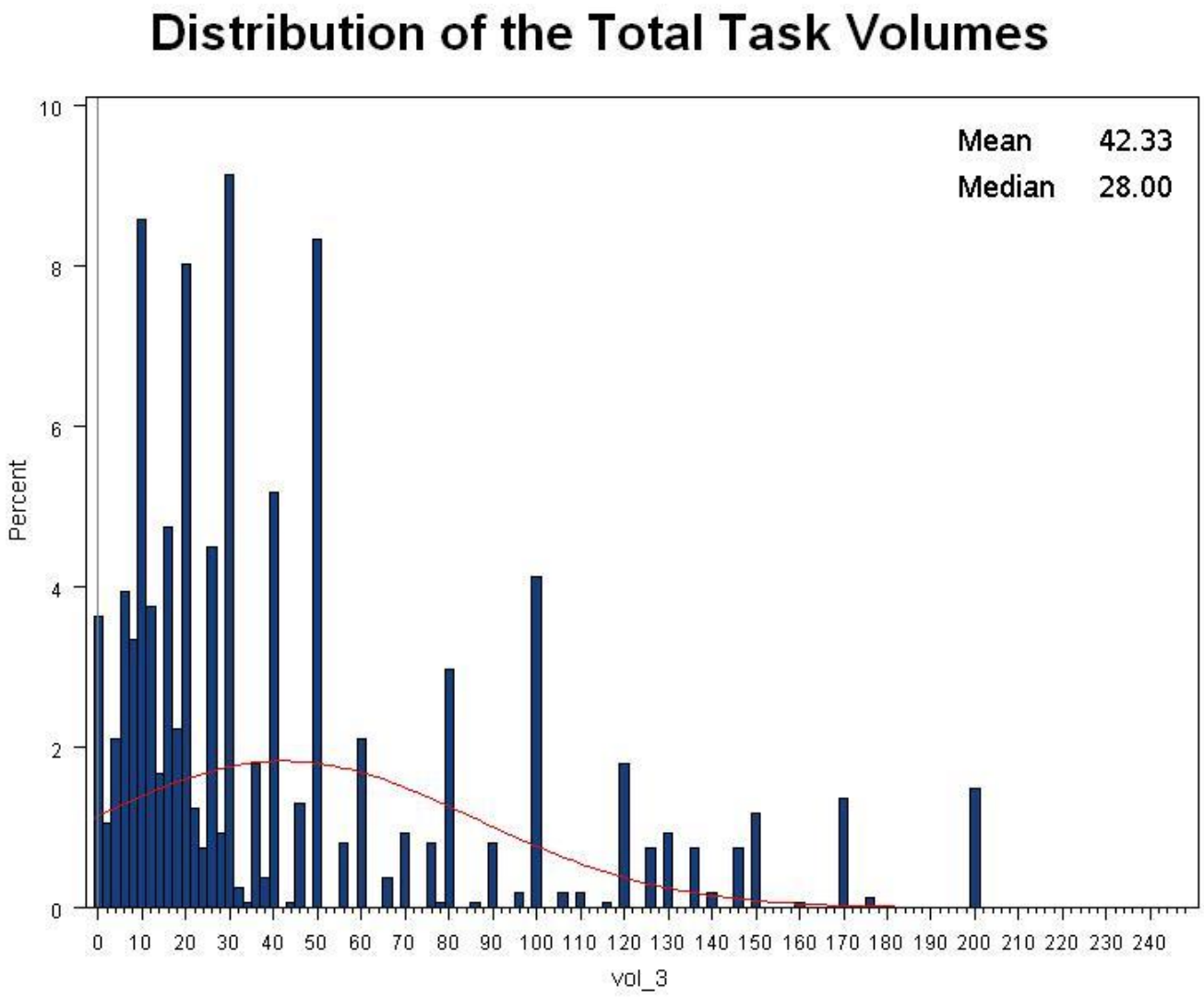


Figure 4: Data set A's distribution of maximum expected volumes

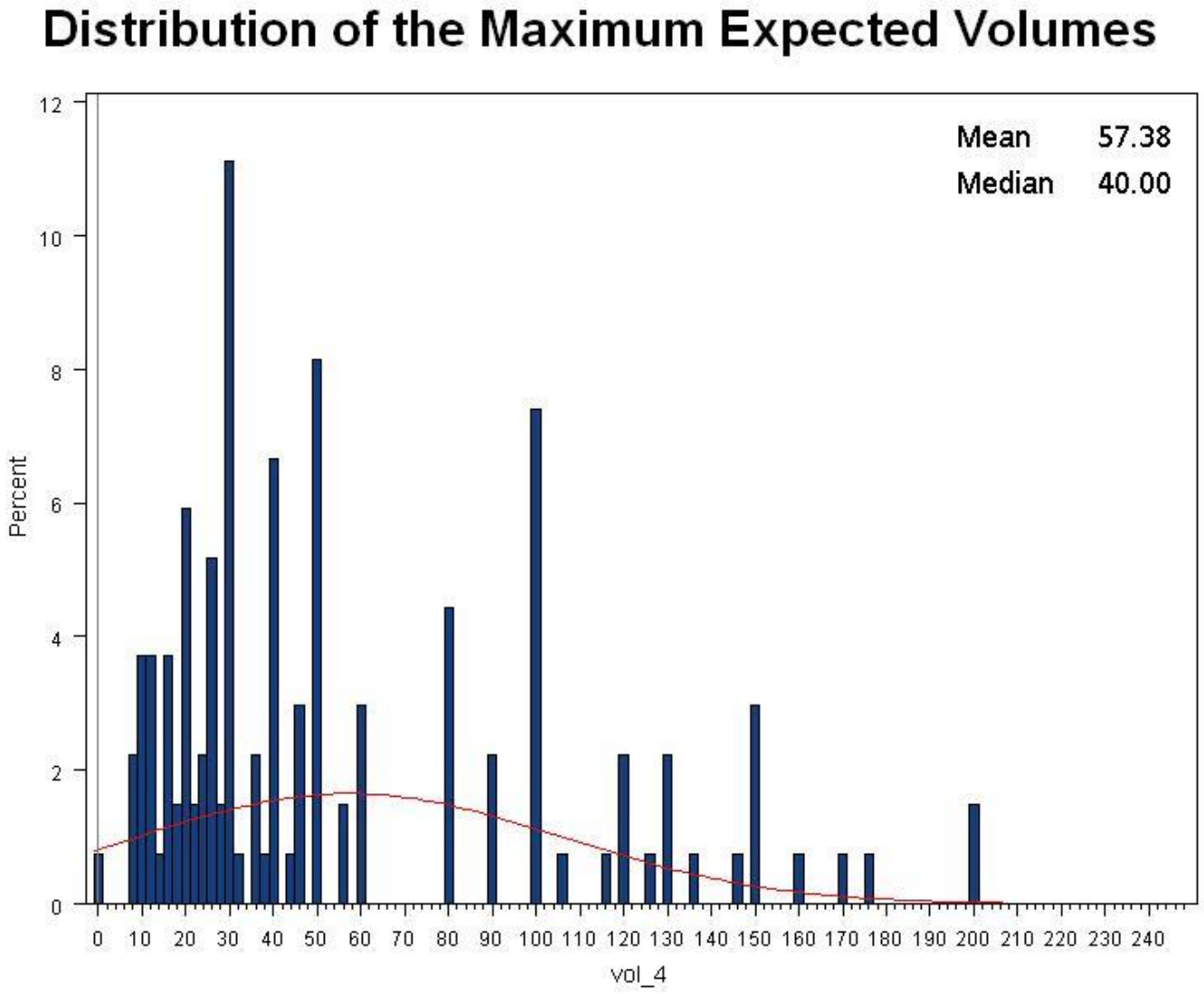


Figure 5: Data set A's actual vs. predicted volumes for the choice modeling approach: holdout data

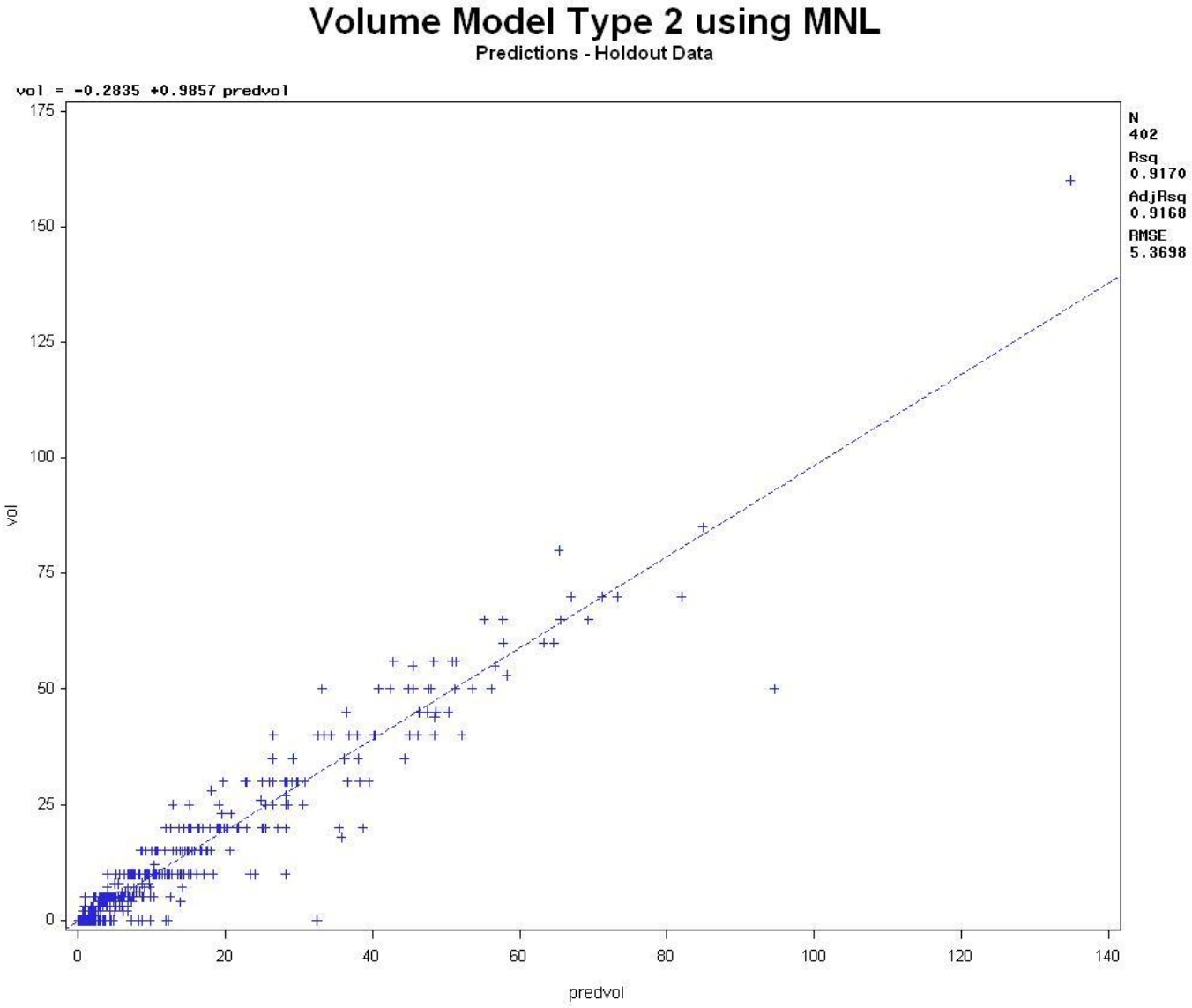


Figure 6: Data set A's actual vs. predicted volumes for the joint D/C approach: holdout data

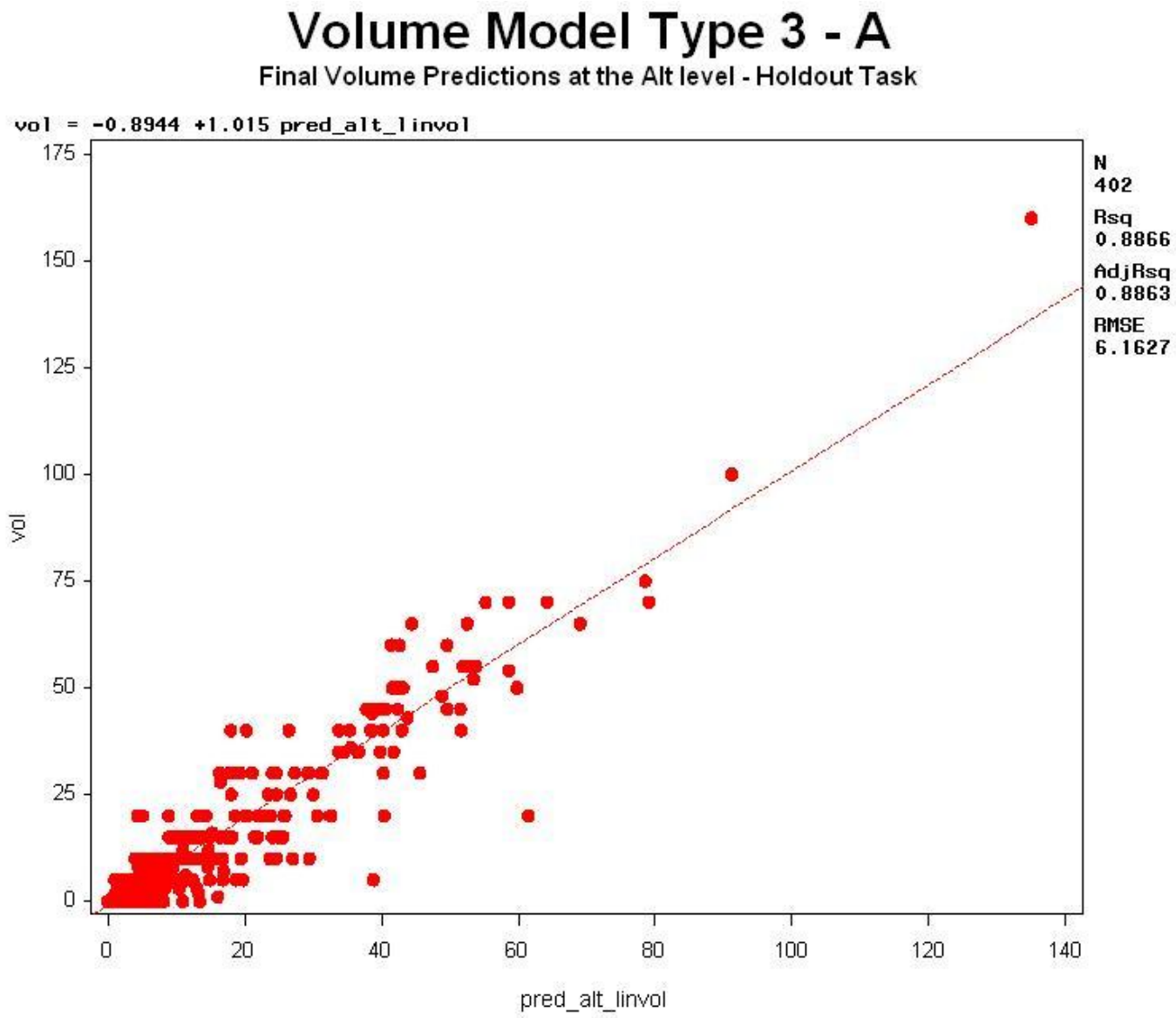


Figure 7: A more complex comparison's actual vs. predicted volumes for the choice modeling approach: holdout data

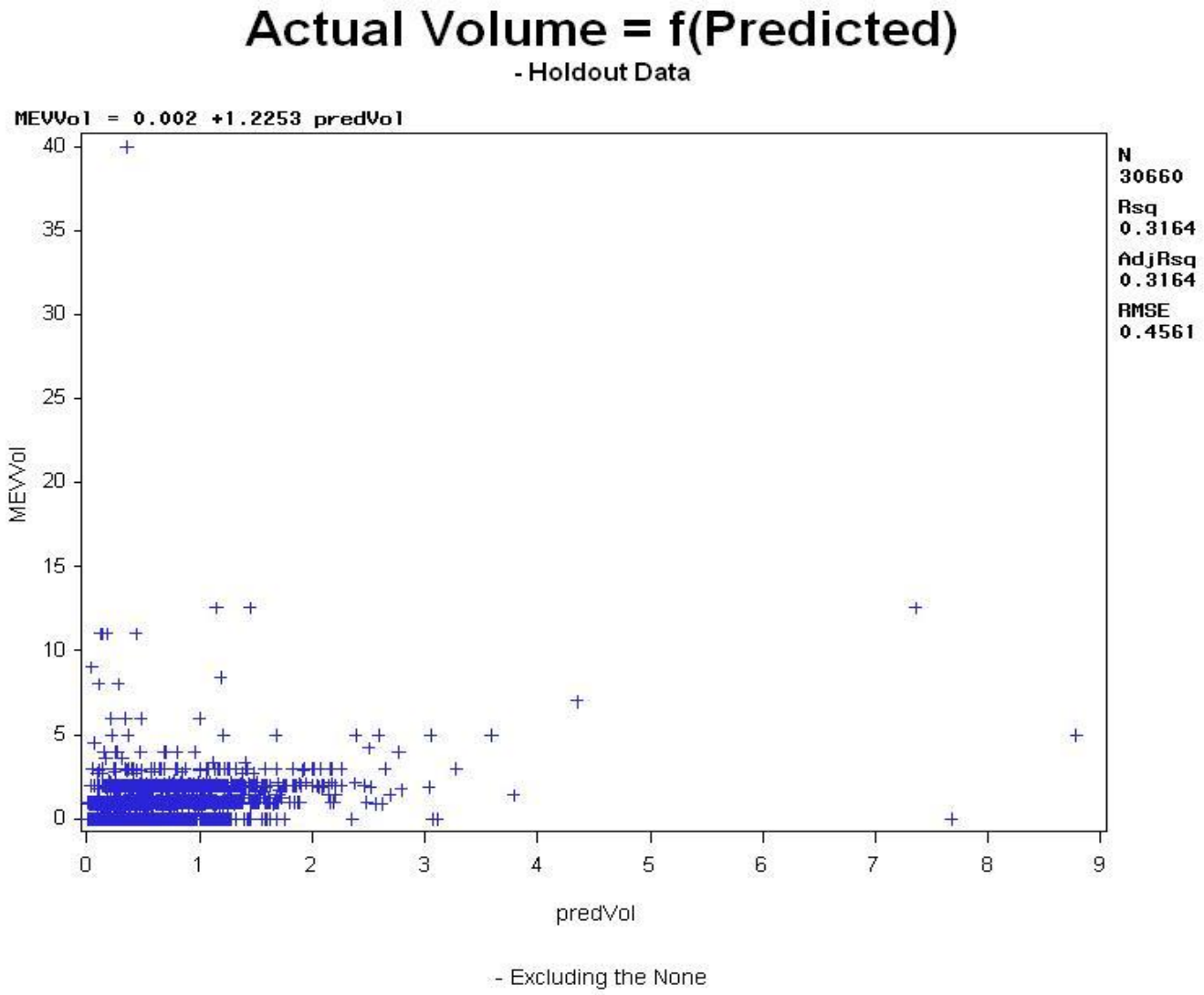


Figure 8: A more complex comparison's actual vs. predicted volumes for the TOTAL VOLUME joint D/C approach: holdout data

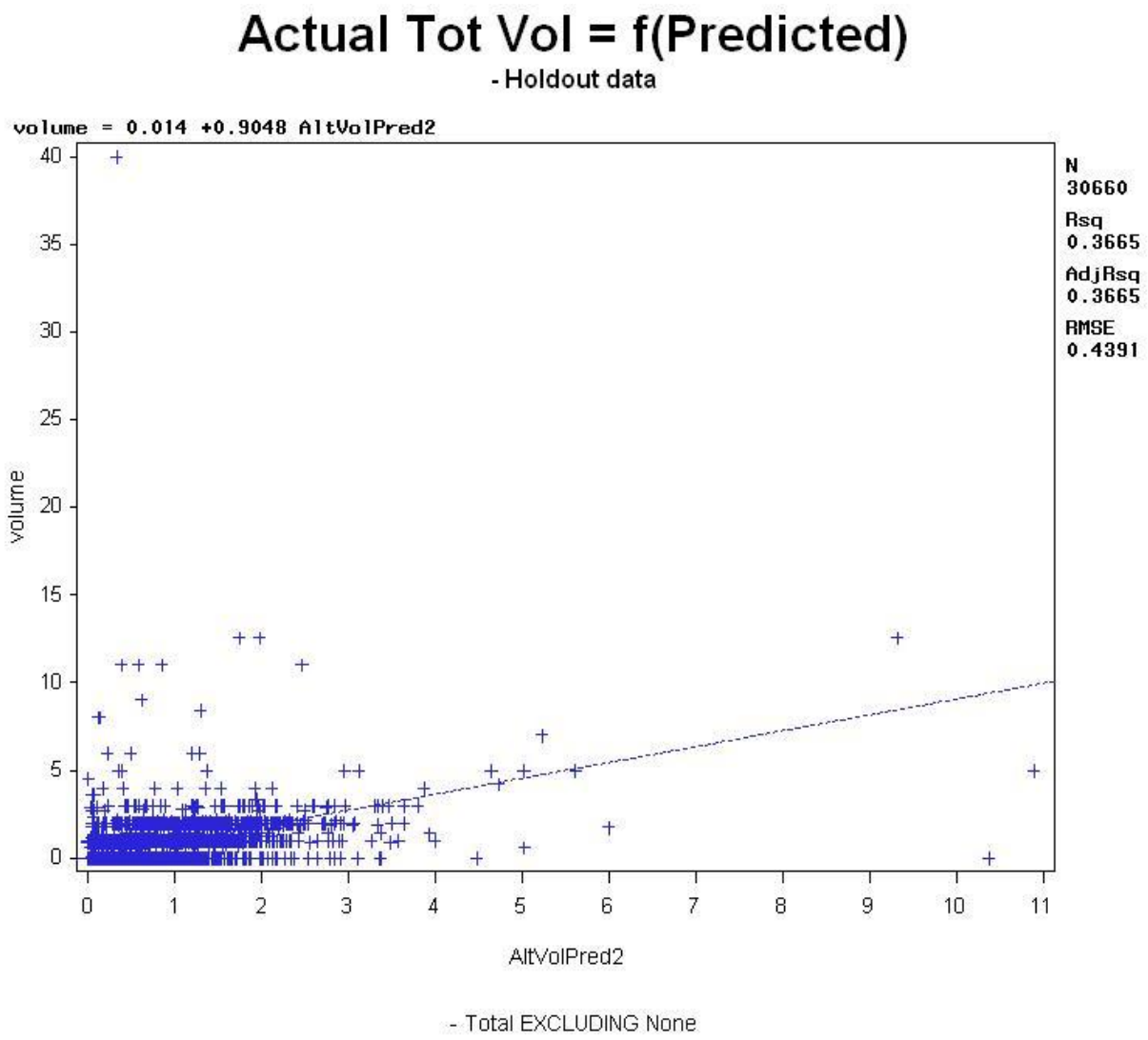
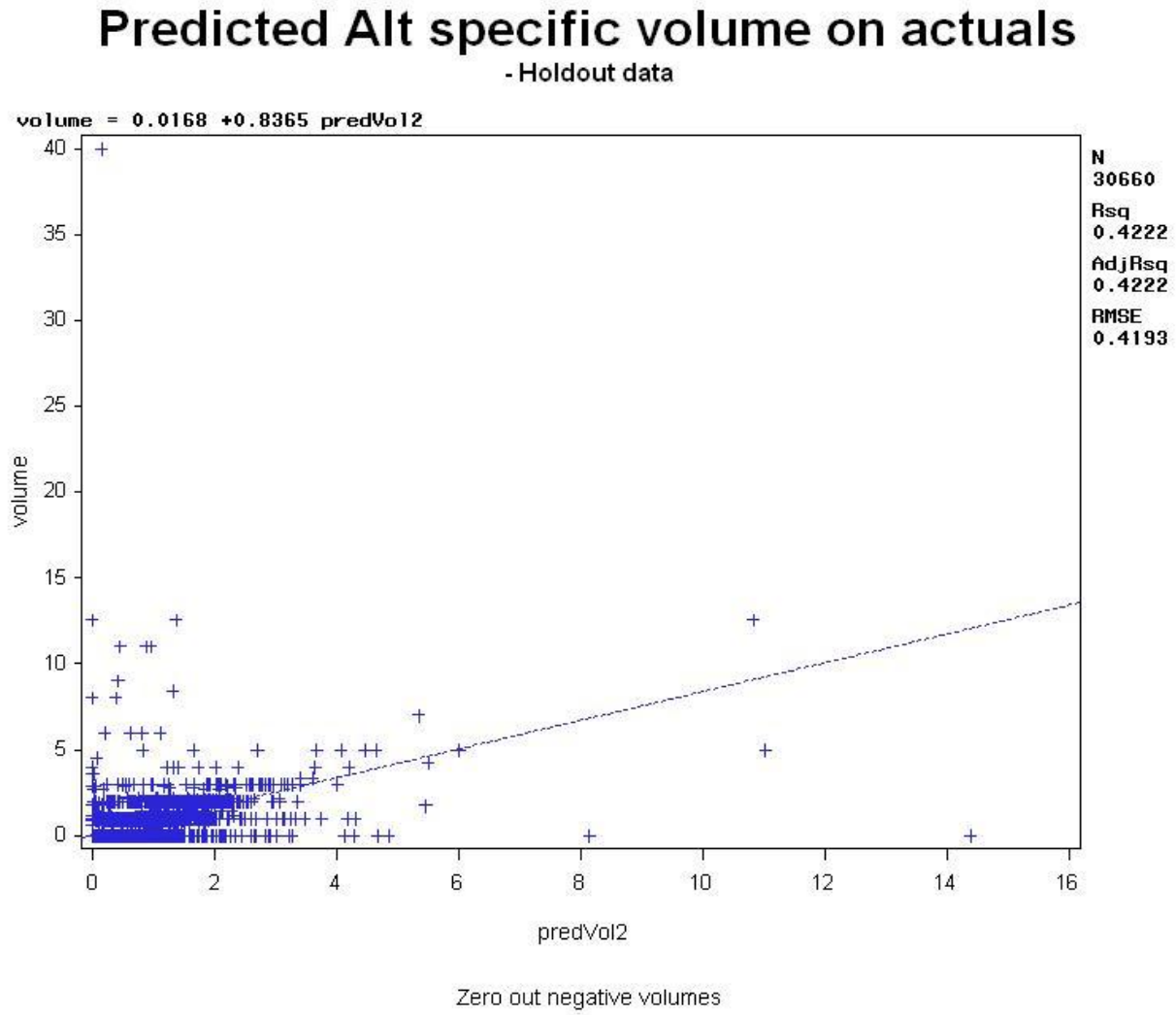


Figure 9: A more complex comparison's actual vs. predicted volumes for the ALT SPECIFIC VOLUME joint D/C approach: holdout data



A Head-to-Head Comparison of the Traditional (Top-Down) Approach to Choice Modeling with a Proposed Bottom-Up Approach

Don Marshall
TVG Marketing Research and Consulting
Siu-Shing Chan
University of Pennsylvania
Joseph Curry
Sawtooth Technologies

Introduction

At the 2009 Sawtooth Software Conference, Jordan Louviere presented data suggesting that the methods used in the design and analysis of choice experiments for the last 25 years were in fact providing biased and misleading information. In his presentation, he summarized some of the research on the design and analysis of choice experiments that had been published over the previous 16 years documenting that with the most common approach adopted, the error term is confounded with the parameter coefficients. Specifically, he asserted that the differences in preferences that we measure among individuals are due more to the way we handle variation in their ability to clearly and consistently express their preferences than to variation in their actual preferences per se.

Louviere then proposed a new way (or more precisely, a move back towards the traditional modeling roots of the field) to estimate choice models (Louviere et al. 2008) that would allow the estimation of an unbiased choice model for each individual respondent which could then be aggregated across all respondents to provide reliable market forecasts. While the most commonly used modeling approach does not take account of differences in error variance between and within individuals, this proposed approach makes assumptions about indirect utility functions, choice processes and error distributions for each individual, potentially providing a more accurate and reliable model for each person separately.

Following some discussion about the implications of the Louviere presentation, and a reanalysis of some old choice data that showed significant differences in the scale factor across respondents when analyzed using latent class analysis (Magidson, 2007), we decided to test the proposed approach using a direct, head-to-head comparison of the proposed new approach with the traditional approach (as exemplified by CBC-HB).

THE TEST

This test involved three arms—one arm for the traditional ('Top-Down') approach as exemplified by CBC-HB, a second arm for the proposed ('Bottom-Up') approach, and a third arm to provide out-of-sample data to determine how well each approach performs when predicting the answers given by individuals not used to estimate the models. The traditional approach is viewed as 'Top-Down' (TD) because it relies on Hierarchical Bayesian (HB) analysis to adjust the results for each respondent based on the aggregate distribution of choices. The proposed 'Bottom-Up' (BU) approach, on the other hand, models each respondent individually, untempered by the choices of other respondents.

While both approaches yield models for each individual respondent, the TD approach incorporates the influence of the aggregate choices in the estimation of the individual's model, while the BU approach relies solely on the data for that respondent.

To ensure that the design and analysis of both approaches was performed in an optimal manner, this test was designed as a collaborative effort. Jordan Louviere and his team (including Richard Carson, Bart Frischknecht and John Rose) from the Center for the Study of Choice (CenSoC) provided guidance and insight in the design and analysis of the 'Bottom-Up' approach. Similarly, Rich Johnson and Bryan Orme of Sawtooth Software provided guidance and insight for the 'Top-Down' arm of the test.

All study participants were involved in the design of both tests (both the subject matter and the attributes and levels to be used in each), as well as agreeing on the test measures to be used when evaluating the results. All fieldwork was performed by Western Wats. This was a highly collaborative effort, with nearly 1,000 emails back and forth between the participants in order to obtain understanding and agreement among all parties.

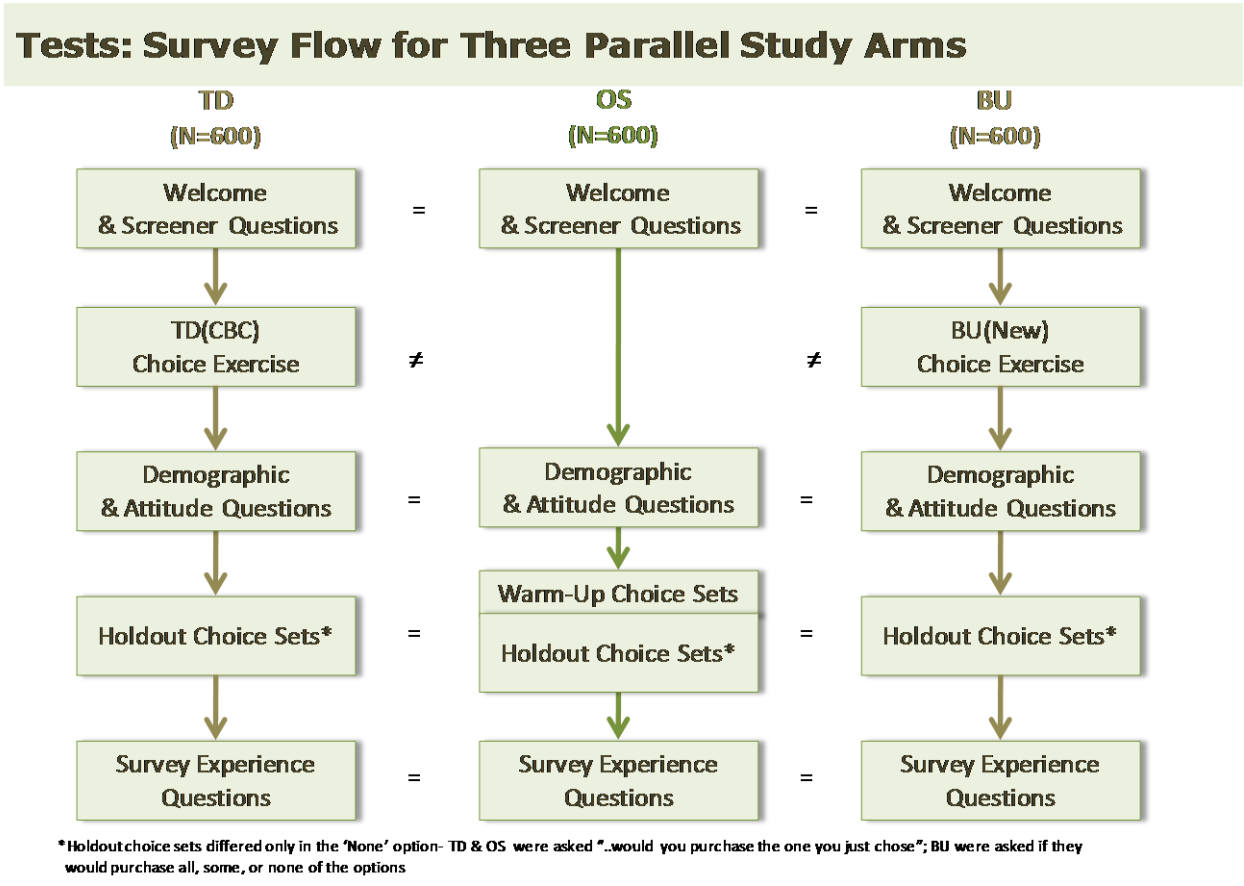
The test included two independent surveys to allow anything learned in the first wave to be incorporated into the second, and to avoid the potential criticism that the results only represented a sample of one (study). The subject for Study 1 was delivered pizza, while Study 2 was on digital cameras. Each wave included 1,800 good respondents (after cleaning out any speeders, straightliners, etc.)—600 who completed the 'Top-Down' survey, 600 who completed the 'Bottom-Up' survey, and 600 who provided the out-of-sample dataset.

Within each wave, the questionnaires were identical except for the choice questions, although the attributes and attribute levels were consistent between the two. The survey started with a welcome screen and a few screening questions. Qualified respondents who were randomly assigned to the TD or BU arms then proceeded to the appropriate choice questions. Following the choice questions, all respondents completed a brief section with demographic and attitudinal questions, followed by five holdout questions. The survey ended with a few questions asking respondents to evaluate their experience when completing the survey.

The TD and BU surveys were identical except for the actual choice questions. The holdout questions for the TD and BU surveys were structured identically to the actual choice questions, but differed in how the 'None' option was included—TD respondents were asked "Would you purchase the one that you just chose", while BU respondents were asked if they would purchase all, some or none of the options in that choice set.

Respondents who were randomly assigned to the out-of-sample arm completed a questionnaire that was identical except that it did not include the choice exercise. Since respondents need to answer a few choice questions to get used to the task, respondents in the out-of-sample arm of the study were asked a few choice questions to allow them to get used to the task prior to answering the actual holdout questions. Figure 1 summarizes the questionnaire flow for each arm of the study.

Figure 1



Those respondents who were randomly assigned to the 'Top-Down' arm were asked the traditional CBC-HB questions that need no further description here. The 'Bottom-Up' approach, however, is a recently proposed technique that is continuing to evolve, and therefore requires some explanation.

As implemented in this study, the 'Bottom-Up' choice questions presented respondents with four designed alternatives. The respondent's task was to review each of these alternatives and indicate which of the four would be their first choice, which would be their last choice, and whether they would actually purchase all of the options in that choice set, some of the options or none of the alternatives presented. All BU respondents saw the same choice sets in the same order. The BU team used the same, fixed design for all respondents. Parameter estimates for a particular design may be biased due to omitted effects, but the choice of one design for all respondents avoids design/preference confounds across people associated with random or blocked designs where each person effectively gets a different design or different part of a large design. The designs utilized by CenSoC try to balance the occurrence of attribute levels across sets, not within sets, and to be as efficient as possible according to a D-efficiency criterion.

The key similarities and differences in design and data collection between the two approaches are summarized in Table 1. The BU approach used a larger design than the TD—20 vs.12 choice sets for the Delivered Pizza study, and 24 vs. 15 for the Digital Camera study. The larger design utilized for the BU approach was not required, but was used purely because the BU team was being "risk averse".

Table 1
Differences in Approach: Data Collection

	Top-Down	**Bottom-Up**
Choice set design:	Respondents see different choice sets	Respondents see the same choice sets
Design:	Balanced overlap design	Balanced overlap design, but with more overlap
Number of choice sets:	Fewer choice sets per respondent (12 and 15 choice sets)	More choice sets per respondent (20 and 24 choice sets)
Choice task:	Less data collected per choice set: -First choice -First choice acceptable or not	More data collected per choice set: -First choice -Last choice -All, some, or none acceptable

For the analysis of the BU data, each respondent was analyzed using weighted least squares regression. The WLS algorithm weights the parameters for respondents based on their consistency. While the TD analysis assumed some order constraints (both waves imposed an order constraint to require lower price to be preferable to higher price and the delivered pizza study also assumed that a faster delivery time should be preferred to a slower delivery), the BU analysis did not employ any constraints. The key analytical differences between the two approaches are summarized in Table 2.

Table 2 Differences in Approach: How Utilities Are Estimated		
	Top-Down	**Bottom-Up**
Use of shared data:	Analysis uses individual and shared data	Analysis uses individual data only
Analysis:	Analyzed using HB, with some order constraints (lower price was assumed to be preferable in both waves; Faster delivery time was assumed to be preferable for the delivered pizza study)	Analyzed using weighted least squares regression with no constraints, and a logit link function

While the TD approach tuned the overall scale factor to minimize the MAE when predicting shares for the in-sample holdout questions, the BU approach did not. Similarities and differences between the two approaches in predicting choice are summarized in Table 3.

Table 3 Differences in Approach: Predicting Choice		
	Top-Down	**Bottom-Up**
Tuning model:	Share of preference with overall scale factor (logit exponent) tuned to best predict in-sample holdout choices)	No tuning of overall scale factor in share of preference
Weighting:	Individual results not weighted	The original WLS estimates are rescaled based on the inverse of the mean square error

The attributes and levels for the Delivered Pizza study (Wave 1) and the Digital Camera study (Wave 2) are shown below.

Attribute	Level 1	Level 2	Level 3	Level 4
Brand	Papa John's	Pizza Hut	Dominos	Little Caesars
Topping	Extra Cheese	Pepperoni	Hawaiian	Veggie Supreme
Crust	Regular Crust	Thick Crust		
Soft Drink	1 liter bottle	2 liter bottle		
Delivery Time	40 minutes	35 minutes	30 minutes	25 minutes
Cost	$14.99	$12.99	$10.99	$8.99

Attribute	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Brand	Canon	Insignia	Kodak	Nikon	Olympus	Sony
Type	Pocket size	Standard				
Megapixels	6	10	12	15		
Optical Zoom	3X	5X	10X	12X		
LCD Display	2”	2.5”	3”			
Image Stabilizer	None	Yes				
Video	None	Standard	High Definition			
Ruggedness	Standard	Waterproof	Shockproof	Waterproof & Shockproof		
Price	$159	$199	$259	$299	$359	

All participants also agreed on the measures that would be used to the performance of the two models. The agreed upon criteria were:

- Calibration fit as measured by Root Likelihood (RLH);
- In-sample holdout hit rate as measured by RLH and hit rate percent;
- In-sample market share prediction accuracy as measured by MAE;
- Out-of-sample market share prediction accuracy as measured by MAE.

General Results

With the larger design used, and the increased number of questions asked for each choice set, it is not surprising to find that the BU approach took significantly longer than the TD. Similarly, the increased respondent burden for BU is reflected in a lower completion rate, particularly for the larger camera study. In terms of the respondents’ evaluation of their experience when completing these questionnaires, the differences, while relatively minor, favor TD. The biggest difference between the respondents’ evaluations of the two approaches is in the respondents level of agreement that the “survey was at times monotonous”, with BU respondents being significantly more likely to agree.

Table 4: General Results				
	Delivered Pizza		Digital Camera	
Measure	TD (CBC & HB)	BU (New & evolving)	TD (CBC & HB)	BU (New & evolving)
Interview length (median, minutes)	8.4	13.4	12.1	20.7
Completion rate	99.2%	97.6%	94.2%	85.8%
Overall experience compared to other internet surveys (5=Far Better)	4.2	4.1	3.8	3.4
	(p<0.001)		(p<0.001)	
Choices seemed realistic (5=Strongly Agree)	4.3	4.3	4.1	4.1
	(p<0.722)		(p<0.105)	
Survey was at times monotonous (5=Strongly Agree)	2.5	2.9	2.9	3.6
	(p<0.001)		(p<0.001)	
Would be interested in taking another survey like this (5=Strongly Agree)	4.5	4.5	4.2	3.8
	(p<0.083)		(p<0.001)	
Survey made it easy to give realistic answers that reflect what I'd do (5=Strongly Agree)	4.2	4.1	4.0	3.8
	(p<0.322)		(p<0.001)	
The way alternative were presented caused me to make careful choices (5=Strongly Agree)	4.0	4.0	4.0	4.0
	(p<0.951)		(p<0.201)	

Turning now to the various goodness-of-fit measures for the two models (see Table 5), we see a somewhat different picture. For this analysis we developed models both with and without 'None' as a predicted outcome. Starting with an assessment of how well each model fits the data, we compared the Root Likelihood (RLH) for the four models. As the data in Table 5 shows, the BU approach had a superior RLH for three out of the four models, with TD having a better RLH for the Digital Camera model 'Without None'. This is as expected since each BU model only needs to fit the data for that respondent, while data from all respondents is involved in estimates for each respondent in the TD case.

Looking next at the in-sample Holdout Hit Rates (using RLH and the percent of correct predictions) shows that the TD approach had a better RLH for three of the models. For one model (Digital Camera without None) the TD RLH was dramatically better—0.26 for TD while the RLH for BU was <0.0001. This very low value may reflect only the fact that a small number of respondents' individual RLH values were very small.) When comparing actual hit rates, TD was better for two of the models, BU for one, while the two approaches were tied for the remaining model.

Comparing MAE for the in-sample Holdout share predictions, TD was better for all four models. Lastly, comparing the MAE for the out-of-sample share predictions shows that TD provides better estimates for both Digital Camera models, while the two approaches are tied for both Pizza models.

Table 5: Model Results TD vs. BU

Measure	Delivered Pizza		Digital Camera	
	TD (CBC & HB)	**BU (New & evolving)**	**TD (CBC & HB)**	**BU (New & evolving)**
Calibration Fit (RLH)				
With None	0.58	0.70	0.58	0.67
Without None	0.63	0.72	0.66	0.35
In-sample Holdout Hit Rate (RLH/%)				
With None	0.37 / 64%	0.35 / 66%	0.25 / 54%	0.20 / 54%
Without None	0.40 / 68%	0.42 / 67%	0.26 / 53%	<0.0001*/ 51%
"Market Share" Prediction in-sample Holdouts (MAE)				
With None	2.6%	3.0%	1.9%	2.4%
Without None	3.2%	4.1%	2.2%	3.6%
"Market Share" Prediction Out of sample Holdouts (MAE)				
With None	3.4%	3.4%	3.1%	3.2%
Without None	4.2%	4.2%	2.8%	3.6%

*Reflects limitation in using geometric mean in computing holdout RLH.

As a last step in this analysis, we wanted to analyze the BU Camera data using Hierarchical Bayes analysis and compare those results with the previous findings. This result is summarized in Table 6. The use of HB on the BU data improved the RLH for the in-sample Holdouts for the model 'With None' from 0.20 to 0.29, but had no effect on the model Without None, while the Hit Rates for both models improved slightly to break the tie in the previous findings.

When looking at the MAE for the in-sample share Predictions, the BU HB model 'With None' was somewhat worse than the BU model, while the use of HB slightly improved the model 'Without None'. Lastly, HB improved the MAE for the out-of-sample share predictions for both models. The BU HB model provided the lowest MAE for any of the out-of-sample models 'With None'.

Table 6: Model Results TD vs. BU vs. BU HB

Measure	Digital Camera		
	TD (CBC & HB)	**BU (New & evolving)**	**BU HB (New & evolving)**
In-sample Holdout Hit Rate (RLH/%)			
With None	0.25 / 54%	0.20 / 54%	0.29 / 55%
Without None	0.26 / 53%	0.26 / 53%	0.26 / 54%
"Market Share" Prediction in-sample Holdouts (MAE)			
With None	1.9%	2.4%	2.9%
Without None	2.2%	3.6%	3.4%
"Market Share" Prediction out of sample Holdouts (MAE)			
With None	3.1%	3.2%	3.0%
Without None	2.8%	3.6%	3.5%

SUMMARY

In summary, the proposed new BU approach has many similarities with the traditional TD approach, but also has some significant differences.

In general, the traditional TD approach requires fewer choice sets, and fewer questions within each choice set, resulting in shorter surveys and higher completion rates.

Data collection is similar for both approaches, but BU collects more data from each respondent (it used more choice scenarios, and asks more questions within each scenario).

Model estimation is different in the two approaches, with TD using Hierarchical Bayes, while BU relies on WLS regression, with the WLS parameters for each respondent being rescaled based on the inverse of their MSE.

Choice prediction is done similarly for the two approaches, but TD tunes the model by adjusting the overall scale factor to minimize the MAE for in-sample holdouts while BU does not.

When comparing the various goodness of fit measures for the two models, this analysis has shown that:

- The proposed BU approach provides the better calibration fit (RLH) for three of the four models. This is as expected since each BU model only needs to fit the data for that respondent, while data from all respondents is involved in estimates for each respondent in the TD case. Since these models are developed to predict market performance of potential new products, their ability to predict out-of-sample responses is a more important criterion in their evaluation.
- The traditional TD approach provides a better hit rate for the in-sample holdouts for two of the four models, BU does better for one, and the two approaches are tied on the remaining model;
- TD provides a better MAE for the in-sample holdout share predictions for all four models;
- For the all important out-of-sample share predictions, TD yielded a better MAE for the more complex camera models, while the two methods were tied for both Pizza models.

Applying HB analysis to the BU camera data improved the BU model performance slightly:

- BU HB provided the best RLH for the in-sample holdouts for one model, while TD and BU HB were tied for the remaining model; BU HB provided the best hit rate for both models;
- The TD model provided better MAEs for both in-sample holdout share predictions;
- For the out-of-sample share predictions, TD provided a lower MAE for one model, while BU HB had a better MAE for the remaining model.

CONCLUSIONS

While the design and analysis criteria for BU continue to evolve (in fact, they evolved in the course of the two waves of this study), this analysis provides no compelling reason to recommend BU over TD at this point. Interview length and completion rates favor TD. Out-of-sample share predictions favor TD for the camera test, while the two approaches are tied for the smaller, less complex pizza test. Refining the BU analysis with HB improves the out-of-sample share predictions, but not enough to recommend using BU in place of TD. As the design and analysis criteria for BU continue to evolve and improve, this may change.

REFERENCES

Islam, Towhidul, Louviere, Jordan & Pihlens, David (2009) — Aggregate choice and individual models: A comparison of top-down and bottom-up approaches, Proceedings, Sawtooth Software Conference, Delray Beach, Florida (March)

Louviere, J.J. & Eagle, T. (2006) — Confound It! That pesky little scale constant messes up our convenient assumptions, Proceedings, 2006 Sawtooth Software Conference, 211-228: Sawtooth Software, Inc., Sequim, WA, USA

Louviere, J.J., Street, D., Burgess, L. Wasi, N., Islam, T. & Marley, A.A.J. (2008) — Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information, The Journal of Choice Modeling, 1, 1, 128-163.

Magidson, J. & Vermunt, J.K. (2007) — Removing the scale factor confound in multinomial logit choice models to obtain better estimates of preference, Proceedings, Sawtooth Software Conference, Santa Rosa, CA (October)

Street, D.J. & Burgess, L. (2007) — The Construction of Optimal Stated Choice Experiments: Theory and Methods. Hoboken, New Jersey: Wiley.

Swait, J. & Louviere, J. J. (1993) — The role of the scale parameter in the estimation and comparison of multinomial logit models, Journal of Marketing Research, 30, 305-314.

HB-CBC, HB-BEST-WORST-CBC OR NO HB AT ALL?

RALPH WIRTH
GFK MARKETING SCIENCES

SUMMARY

Choice-Based Conjoint Analysis (CBC) is currently the most popular Conjoint approach (Sawtooth Software 2010). Nevertheless, concerns have been discussed in the research community regarding (1) the quality of Hierarchical Bayes (HB) parameter estimation in CBC studies (HB-CBC) when data conditions are particularly challenging and (2) the ability of these HB choice algorithms to yield correct results when individual error variances are not equal across the sample. This paper introduces two alternative approaches to CBC that were developed in order to deal with these potential problems: The Louviere et al. (2008) approach does not rely on HB-estimation, thus circumventing potential problems with individual-specific error variances. Best-Worst-CBC (BW-CBC, Wirth 2010) aims at extracting more preference information from CBC-exercises by asking for both the best and the worst option in each choice task, thus improving data conditions for HB estimation.

Since little research has been done on how to correctly model Best-Worst choices and on how to set up an HB-model for estimating parameters based on BW-CBC (called HB-BW-CBC in the following), different probabilistic BW-choice models derived by Marley and Louviere (2005) are introduced and assessed based on their fit to empirical data sets. A subsequent comprehensive simulation-based model comparison reveals that both HB-CBC and HB-BW-CBC work very well - even under sparse data conditions and when individual error variances differ a lot. The Louviere et al. approach yields satisfactory results in more comfortable data situations and turns out to be a simple approach that is worth considering when the focus is on share prediction rather than on the prediction of individual choices. Finally, an empirical comparison of HB-CBC and HB-BW-CBC reveals that the superiority of the Best-Worst approach, which is observed in the simulation study, can also be confirmed on real data.

INTRODUCTION: HB-CBC AND ITS (ASSUMED) PROBLEMS

Over the last decade, Choice-Based Conjoint-Analysis (CBC) has become the most popular Conjoint approach worldwide. One main reason for that is the increasing computer power which now enables researchers to utilize complex Hierarchical Bayes (HB) algorithms for estimating individual utility parameters in spite of the sparse nature of choice data. While many empirical and simulation-based research studies show that the HB-CBC approach generally yields good results (see Moore et al. 1998, Teichert 2001, Gensler 2003, Moore 2004, Pinnell 2004, Hillig 2006), two main concerns are still discussed in the research community:

1. There is evidence that the goodness of the HB-estimation decreases dramatically when data conditions are too challenging. For example, Andrews et al. (2002a) conclude in their influential article on the results of a comprehensive simulation study they conducted:
   "The most important new finding of this study […] is the poor performance of the [..]

logit model with HB estimation when parameters are poorly identified at the individual level."

2. Most standard HB-choice models implicitly assume that respondents are homogeneous with regard to their error variances. Some researchers, like e.g. Louviere and Eagle (2006) emphasize that the violation of this assumption may lead to erroneous HB-estimations:
"[…] it is highly likely that [these] models are biased and misleading".

It has already been shown that individual error variances are unlikely to be constant across the sample, as they may well depend on individual characteristics such as education or age (Louviere et al. 2002). Furthermore, the data conditions that were investigated by Andrews et al. (2002) are rather the rule than the exception in many commercial CBC studies. Hence, both potential problems mentioned above are of high relevance to the market research community. Each of the two approaches that will be introduced in the following chapter addresses one of these concerns: The main objective of ***Best-Worst-CBC*** (Wirth 2010) is the efficient gathering of additional preference information so that data conditions for HB estimation improve. In contrast, the ***Louviere et al. approach*** (Louviere et al. 2008) also uses additional preference information (rankings) and does not rely on HB estimation at all, thus circumventing the second potential problem mentioned above.

ALTERNATIVE CBC APPROACHES

Best-Worst Choice-Based Conjoint Analysis

Clearly, the best strategy to deal with challenging data situations would be a modification of the design settings. However, often neither reducing the number of parameters (i.e. attributes and levels), nor increasing the number of choice tasks are practical options. Therefore, it seems reasonable to try to extract more information about respondents' preferences from each single choice task. One possible approach for that is the use of Best-Worst questions. *Figure 1* shows an example of a Best-Worst choice task in the context of BW-CBC[1]: Instead of simply asking for the best alternative within each choice task, you ask for the best ***and the worst*** alternative. The none option can easily be taken into account, e.g. by following a Dual Response None approach (Brazell et al. 2006; Diener et al. 2006).

[1] Note that Best-Worst CBC is not the same as Best-Worst Conjoint, which was introduced by Swait, Louviere and Anderson (1995, see also Chrzan and Fellerman 1997). Whereas in BW-Conjoint respondents have to choose the best and the worst *attribute level within each shown concept*, the task in BW-CBC is to choose the best and the worst fully specified *concept* from a set of alternatives. A more comprehensive introduction to different kinds of Best-Worst choices can be found in Marley (2010).

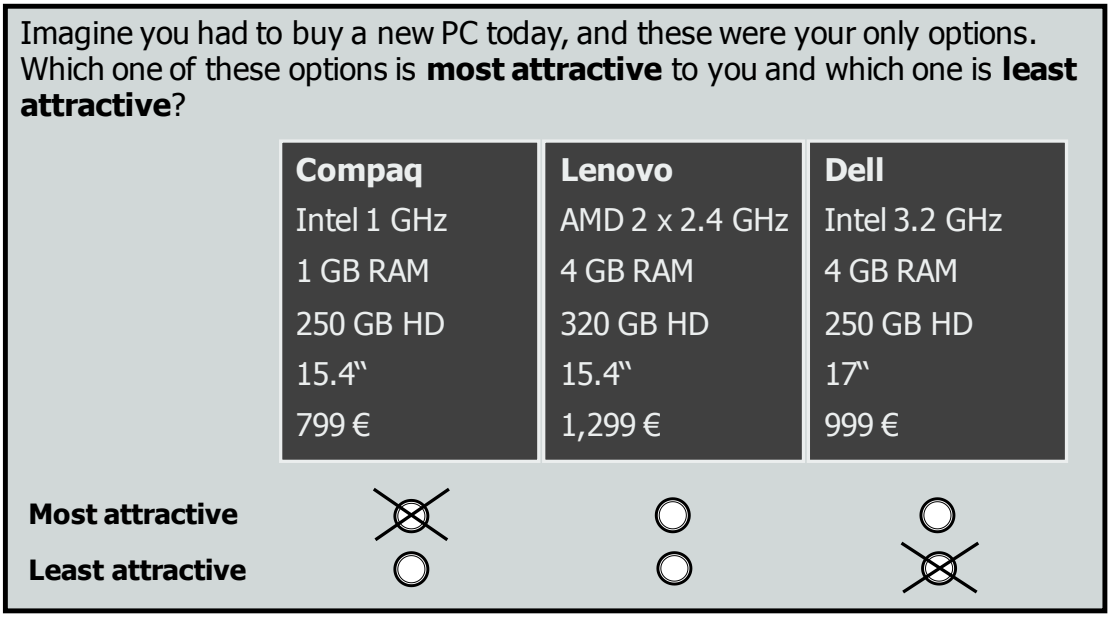


Figure 1
Best-Worst Choice Task

The theoretical advantages of this kind of Best-Worst questions can easily be illustrated based on the example of one respondent facing a choice task with four options {A,B,C,D}. Suppose the respondent chooses alternative A as best. Based on this information, it can be inferred that A≻B, A≻C and A≻D, with ≻ meaning "is preferred to". Now suppose that it is also known that D is the worst alternative for this respondent. Based on this information it can be further inferred that B≻D and C≻D. Hence, the only pair for which the preference order cannot be derived is {B,C}. Since it has already been shown that a second choice within one choice task is made a lot quicker than the first choice (Johnson, Orme 1996), Best-Worst style questions appear to be a very efficient way to extract more information about peoples' preferences from a choice experiment.

Of course, the statistical model for parameter estimation has to be adapted to the different choice situation in BW-CBC. Marley and Louviere (2005) derive different classes of suitable probabilistic Best-Worst choice models. The model classes are very flexible and different assumptions lead to different concrete choice models. Four of these concrete Best-Worst choice models will be briefly introduced in the following sections.[2] The following notation will be used:

- $B_C(j')$ denotes the probability that alternative j' is chosen as best option in choice set C,
- $W_C(j'')$ denotes the probability that alternative j'' is chosen as worst option in C,
- $BW_C(j',j'')$ denotes the probability that j' is chosen as best option and j'' is chosen as worst option $(j' \neq j'')$ in choice set C and
- J is the number of alternatives in choice set C.

[2] For more details and derivations see Marley, Louviere (2005) or Wirth (2010).

For the sake of simplicity, indices referring to individuals and choice tasks are omitted.

The MaxDiff Model

The MaxDiff model is a relatively well-known approach for modeling Best-Worst choices. Based on Random Utility Theory (RUT) and given certain assumptions about the observable and the random component of utility, one can derive that

$$BW_C(j',j'') = \frac{\exp(V_{j'} - V_{j''})}{\sum_{\substack{r,s \in C \\ r \neq s}} \exp(V_r - V_s)} \quad (j' \neq j''),$$

with

V_r being the observable utility of alternative r.

Usually, V_r is specified linear in parameters, i.e.

$$V_r = \boldsymbol{\beta}' \mathbf{x}_r$$

with

$\boldsymbol{\beta}$ being the vector of parameters (partworth utilities and/or linear parameters) and

$\mathbf{x}_r$ being the design vector that describes alternative r in terms of the attributes and levels.

The Consistent Extreme Value Random Utility Model

Slightly different assumptions within the framework of RUT lead to the Consistent Extreme Value Random Utility Model. Following this model, the probability of a particular pair of alternatives j' and $j'' \in C$ $(j' \neq j'')$ being chosen as best and worst can be calculated as follows:

$$BW_C(j',j'') = B_C(j') \sum_{\eta \in R(C-\{j',j''\})} B_{C-\{j'\}}(\eta_2) \dots B_{\{\eta_{J-1},j''\}}(\eta_{J-1}),$$

with

$C - \{j',j''\}$ denoting the choice set C without the alternatives j' and j'',

$R(C - \{j',j''\})$ denoting the set of rank orders of $C - \{j',j''\}$,

$\eta = \eta_2 \eta_3 \dots \eta_{J-1}$ being one concrete rank order from the set $R(C - \{j',j''\})$, with the elements $\eta_2, \eta_3, \dots \eta_{J-1}$ and

$B_{C-\{j'\}}(\eta_2)$ to $B_{\{\eta_{J-1},j''\}}(\eta_{J-1})$ all being calculated using the standard Multinomial Logit (MNL) model.

The Biased MaxDiff Model

The Biased MaxDiff Model can be derived based on certain assumptions about people's choice behavior. It converges to the standard MaxDiff model for large choice set sizes J, as can be seen when looking at the formal representation of the Best-Worst choice probabilities:

$$BW_C(j',j'') = \frac{\exp(V_{j'} - V_{j''}) + \frac{1}{J-1}}{\sum_{\substack{r,s\in C \\ r\neq s}} \left[\exp(V_r - V_s) + \frac{1}{J-1}\right]}$$

The Concordant Best-Worst Choice Model

The most complex BW-choice model among the four approaches presented here is the Concordant Best-Worst Choice Model. It is derived based on the assumption that the best and the worst alternative are chosen sequentially in each choice set, and that the order of these choices (i.e. first best, then worst or first worst, then best) does not affect the BW-choice probabilities. Formally, this means that:

$$\begin{aligned} BW_C(j',j'') &= B_C(j')W_{C-\{j'\}}(j'') \\ &= W_C(j'')B_{C-\{j''\}}(j') \,. \end{aligned}$$

Marley (1968, Theorem 7) shows that, under weak technical assumptions, this property only holds, if and only if the Best and Worst choice probabilities are generally calculated as follows (see Marley, Louviere 2005):

$$B_Y(x) = \frac{\sum_{\rho\in R(Y-\{x\})} b_Y(x\rho)}{\sum_{\eta\in R(Y)} b_Y(\eta)} \qquad \text{and} \qquad W_Y(y) = \frac{\sum_{\rho\in R(Y-\{y\})} w_Y(y\rho)}{\sum_{\eta\in R(Y)} w_Y(\eta)},$$

with

$B_Y(x)$ denoting the probability of an alternative x being chosen as best in a choice set Y,

$W_Y(y)$ denoting the probability of an alternative y being chosen as worst in a choice set Y,

$R(Y)$ denoting the set of rank orders of Y,

$R(Y - \{x\})$ (respectively, $R(Y - \{y\})$) denoting the set of rank orders of Y without alternative x (respectively, y).

With choice set Y being of size Z,

$\eta = \eta_1\eta_2 \ldots \eta_Z$ is a concrete rank order of Y, with elements $\eta_1, \eta_2, \ldots \eta_Z$.

η can also be denoted by $\eta = \eta_1\rho$, with

$\rho = \eta_2 \ldots \eta_Z \in R(Y - \{\eta_1\})$.

Finally, $b_Y(\eta)$ and $w_Y(\eta)$ are defined as

$$b_Y(\eta) = \prod_{1 \le i < j \le Z} b(\eta_i, \eta_j) \quad \text{and} \quad w_Y(\eta) = \prod_{1 \le i < j \le Z} w(\eta_i, \eta_j) \ ,$$

with

$b(\eta_i, \eta_j)$ denoting the binary choice probability of choosing η_i as best alternative in the set $\{\eta_i, \eta_j\}$, and

$w(\eta_i, \eta_j)$ denoting the binary choice probability of choosing η_i as the worst alternative in the set $\{\eta_i, \eta_j\}$.

The property that the BW-choice probabilities of the Concordant BW Choice Model are not affected by the order of best and worst choices holds irrespective of the representations of the binary choice probabilities, as long as they satisfy the assumption that

$$b(x,y) = w(y,x).$$

Therefore, this model should be seen as a model class rather than a concrete model. All results related to the Concordant BW Choice Model that are presented in this paper are based on a Binary Logit representation of the binary choice probabilities.

Parameter Estimation

Parameter estimation for BW-CBC can be done using an HB framework. The four probabilistic BW choice models presented above can be used to specify the likelihood function of the HB model. Apart from the likelihood function, the four possible HB models for parameter estimation do not have to differ. For example, the same standard conjugate priors can be used for the four HB models. When doing so, the MCMC algorithm for sampling from the posterior distribution differs from the standard HB-CBC algorithm only in terms of the Metropolis-Hastings step, in which the likelihood is evaluated using one of the probabilistic BW choice models instead of using the standard MNL model. Detailed information about the hierarchical

model and the MCMC sampler that were used for the estimations presented in this paper can be found in Wirth (2010, pp. 149-153).

The Louviere et al. Approach

Other than both HB-CBC and HB-BW-CBC, the Louviere et al. approach aims at a purely individual estimation of utility parameters:

> *"Other approaches to estimating model parameters for single persons are based on continuous or finite distribution models […]. These approaches model individuals indirectly […], whereas our focus […] is on modeling individuals directly, not modeling them indirectly based on assumptions about preference distributions across samples of people"(Louviere et al. 2008, p. 129).*

In order to get enough information for such a direct modeling of individual preferences, Louviere et al. also use Best-Worst questions, but they go one step further than researchers do in the case of BW-CBC: They suggest to ask ***repeatedly*** for the best and the worst option in each choice task until the full preference order is obtained. A typical Louviere et al. task is illustrated in *Figure 2*. Based on the choices in this task, the full preference order A≻B≻D≻C can be inferred.

	Alternative A	Alternative B	Alternative C	Alternative D
Price	500 €	750 €	800 €	400 €
Brand	Brand X	Brand Y	Brand Z	Brand W
Color	green	blue	yellow	red
Which alternative do you prefer most?	☒	☐	☐	☐
Which alternative do you prefer least?	☐	☐	☒	☐
Which of the remaining two alternatives do you prefer?	☐	☒	☐	☐

Figure 2
Louviere et al. Choice Task

The derived preference orders are then used to estimate individual utility parameters. Louviere et al. (2008) suggest different possible approaches for doing so. One easy way, which is also followed throughout this paper, is the following:

1. Expand the preference ranking into choice sets

Consider all possible non-empty, non-singleton subsets of the alternatives in the current choice set and infer the implicit "winners" in these subsets from the derived preference ranking.

For example, the preference order A≻B≻D≻C would be expanded into the following eleven non-empty, non-singleton choice sets (implicitly chosen options are **bold**):

- {**A**,B}, {**A**,C}, {**A**,D}, {**B**,C}, {**B**,D}, {**C**,D}
- {**A**,B,C}, {**A**,B,D}, {**A**,C,D}, {**B**,C,D}
- {**A**,B,C,D}

2. Estimate individual utility parameters based on the large amount of expanded choice tasks using e.g. the MNL model and standard Maximum Likelihood techniques for fitting the Best choices.

Following these explanations, it becomes obvious that the Louviere et al. approach is much simpler and much more easy to understand for practitioners than HB approaches, which are often seen as black boxes due to their complexity, as pointed out e.g. by Poynter (2006, p. 1):

"A potential problem introduced by Hierarchical Bayes is the loss in transparency. Most researchers will not be able to understand the inner workings of the Hierarchical Bayes, something which will make many uncomfortable."

Therefore, it will be interesting to see if – or in which situations – an approach as simple as the Louviere et al. approach can compete with the complex HB approaches. Among other things, this question will be addressed in the research study that is going to be presented in the following.

RESEARCH STUDY

Overview of Research Approach

The flow of the research project is illustrated in *Figure 3*. It comprises two successive stages. The main focus of the research study is on a systematic, simulation-based comparison of HB-CBC, HB-BW-CBC and the Louviere et al. approach (stage 2). However, as explained above, there are various probabilistic Best-Worst choice models that can be used for specifying the likelihood function of the HB-BW-CBC model. Since, due to time constraints, taking into account all four probabilistic BW choice models presented above in the simulation study is not a feasible option, it is necessary to choose one of the models in advance. In order to base this selection on empirical insights, a comparison of the alternative models with regard to their fit to three data sets is conducted (stage 1 of the research project).

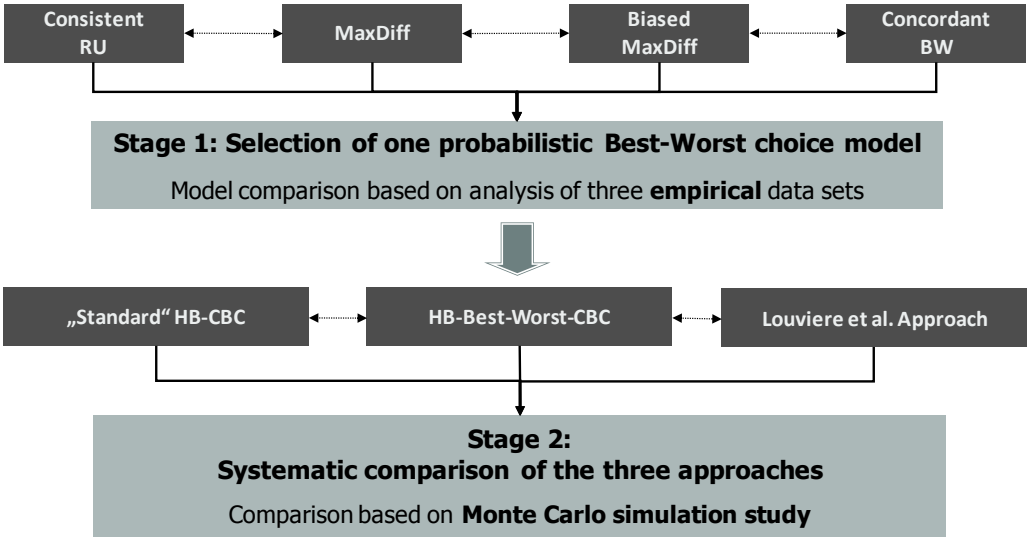


Figure 3
Overview of the Research Project

Stage 1: Empirical Comparison of Probabilistic Best-Worst Choice Models

General Approach, Data Sets and Performance Measures

Figure 4 illustrates the general approach to the empirical comparison.

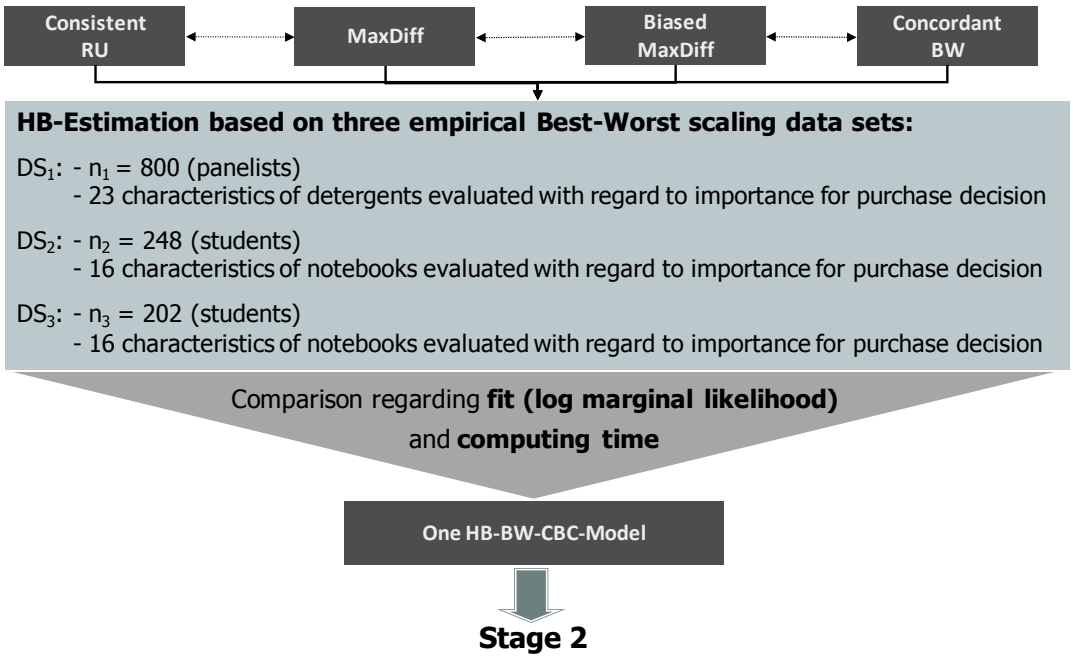


Figure 4
Stage 1: Empirical Selection of One Probabilistic Best-Worst Choice Model

First, four HB-BW-CBC models are specified that differ only in terms of their likelihood functions. That is, these likelihood functions are specified using the Consistent Extreme Value Random Utility Model (Consistent RU), the MaxDiff Model, the Biased MaxDiff Model and the

Concordant Best-Worst Choice Model respectively. Then, the four HB-BW-CBC models are used to estimate parameters based on three different Best-Worst scaling data sets whose characteristics are presented in *Figure 4*, too.[3]

Since the objective of this stage of the research project is to identify the model that matches the true data-generating process best, the four alternatives are compared in terms of fit to the data sets. A popular measure in Bayesian statistics for model comparisons based on fit is the marginal likelihood,[4] which describes the likelihood of the observed data given a specific model (Congdon 2003, p. 470) and which is defined as

$$p(\mathbf{y}|M_g) = \int \dots \int p(\mathbf{y}|\boldsymbol{\theta}, M_g)\, p(\boldsymbol{\theta}|M_g) d\boldsymbol{\theta},$$

with

$p(\mathbf{y}|M_g)$ denoting the marginal likelihood,

$\mathbf{y}$ denoting the data,

θ denoting the model parameters, and

M_g denoting the g-th model (g=1,…G).

As the required integral is usually not solvable analytically, it is common to use numerical approximations for calculating the marginal likelihood. According to Newton and Raftery (1994) the harmonic mean of the likelihood values evaluated at the MCMC parameter draws (of course, after discarding burn-in draws) approximates the marginal likelihood. That means that the following expression can be used for calculating the approximate marginal likelihood:

$$\hat{p}(\mathbf{y}|M_g) = \left(\frac{1}{T-B} \sum_{t=B+1}^{T} \frac{1}{p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_t, M_g)} \right)^{-1},$$

with

$\hat{p}(\mathbf{y}|M_g)$ denoting the approximation of the marginal likelihood,

T denoting the total number of MCMC-draws

B denoting the number of burn-in draws, and

$p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_t, M_g)$ denoting the likelihood evaluated at the t-th MCMC-draw of the parameters $\boldsymbol{\theta}$ for model M_g.

[3] Of course, the same (diffuse) prior settings are used in the four HB models (see Wirth 2010, p. 165 for details).

[4] Fit measures taking into account the number of parameters (like e.g. AIC, BIC) are not necessary for this model comparison, as all probabilistic BW choice models result in the same number of parameters.

While the Newton-Raftery approximation is widely used, it has some disadvantages. Above all, the calculated numerical values tend to be unstable, especially if the data is not informative about the parameters and diffuse priors are used (Rossi et al. 2005, p. 168 and pp. 173-177). An additional examination of the model-specific sequence plots of the log-likelihood over the MCMC-draws helps to assess whether differences in the marginal likelihood for different models are really relevant or not (Rossi et al. 2005, p. 175). For example, based on the illustrative plot in *Figure 5* one could infer that Model 2 fits the data better than Model 1, as the respective log-likelihood values are consistently lower.

All in all, the fit of the four probabilistic BW choice models to the empirical data sets is compared using

- the log marginal likelihood (LML)[5], calculated using the Newton Raftery approximation and
- the sequence plots of the log-likelihood values.

As the feasibility of the Monte Carlo simulation study depends on the time requirements of the HB approaches taken into account, the computing times are considered in the model comparison as well.

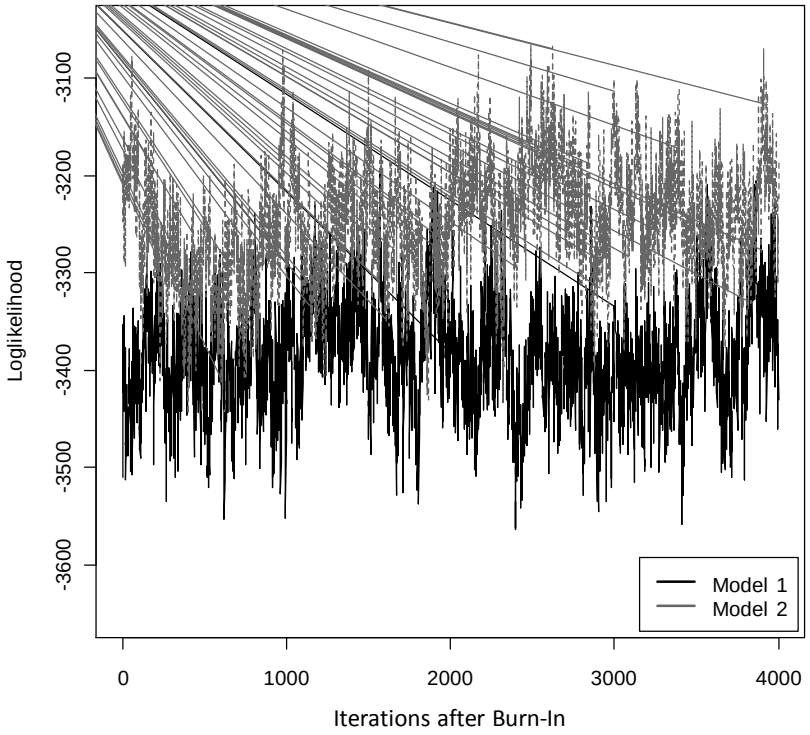


Figure 5
Model Comparison Using Sequence Plots of Log-Likelihood Values

[5] It is common to use the log of the marginal likelihood instead of the marginal likelihood directly for model comparison.

RESULTS

Tables 1, 2 and 3 show the estimated LML-values and the corresponding ranks for the four probabilistic BW choice models based on data sets 1, 2 and 3 respectively. When looking at the ranks, it becomes obvious that there is no empirical evidence for the consistent superiority of any one of the four models: Depending on the data set, different models fit the data best.

	Rank (LML)	LML
Consistent RU	**4**	-3563
MaxDiff	**3**	-3556
Biased MaxDiff	**1**	-3536
Concordant BW	**2**	-3540

Table 1
Log Marginal Likelihood (Data Set 1)

	Rank (LML)	LML
Consistent RU	**3**	-3175
MaxDiff	**2**	-3164
Biased MaxDiff	**4**	-3199
Concordant BW	**1**	-3139

Table 2
Log Marginal Likelihood (Data Set 2)

	Rank (LML)	LML
Consistent RU	**1**	-17682
MaxDiff	**3**	-17914
Biased MaxDiff	**2**	-17897
Concordant BW	**--**[6]	--

Table 3
Log Marginal Likelihood (Data Set 3)

Furthermore, the analysis of the ***sequence plots*** leads to the conclusion that the observed differences in the log marginal likelihood seem to be only marginal. For example, *Figure 6* shows the log-likelihood plot for the four probabilistic BW choice models based on the second data set. Obviously, it is very hard to identify significant differences between the four models.

[6] The estimation could not be carried out due to system requirements that could not be satisfied by the used computer (Intel Dual Core E8400 processor, 8 GB RAM).

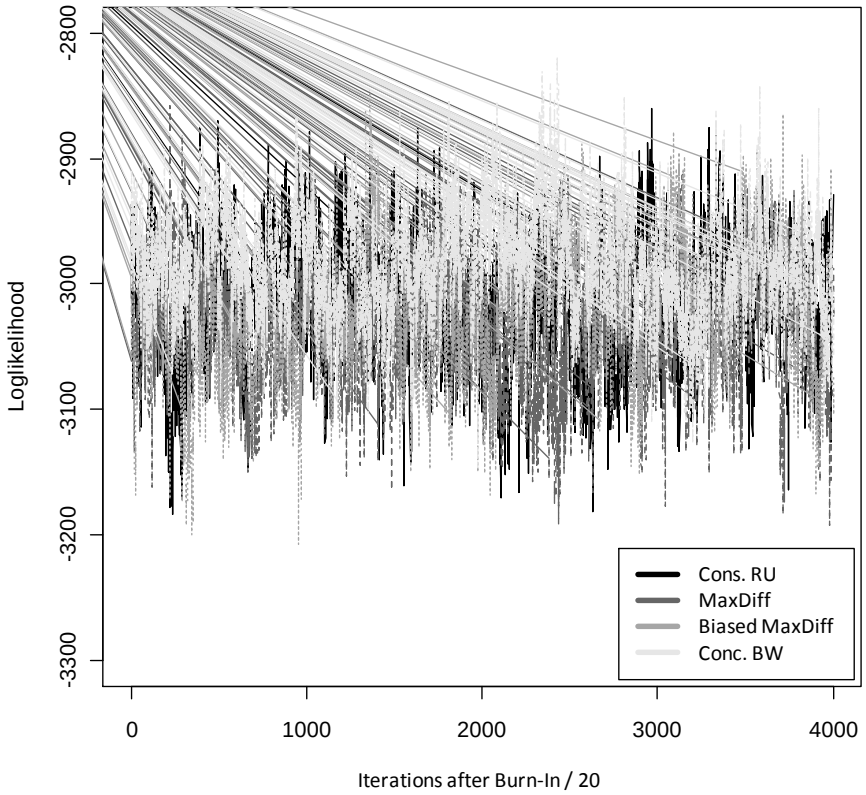


Figure 6
Sequence plot of log-likelihood values (Data Set 2)

Based on these findings, choosing one clear "winner model" with regard to fit to empirical data is difficult. However, when additionally taking into account time requirements of the different models, the picture becomes clearer. The differences in computing time, summarized in *Table 4*, are striking.[7] For example, while HB parameter estimation using the MaxDiff model takes about 2 hours based on the first two rather small data sets, the Concordant BW choice model leads to computation times of about 25 hours in these situations. Even worse, parameter estimation based on the large data set 3 was not possible at all when using the Concordant BW model. Of course, these huge differences can probably be decreased by implementing the HB algorithm in another programming language.[8] In addition, the time requirements of the MaxDiff Model and the remaining two alternatives (i.e. the Consistent RU Model and above all the Biased MaxDiff Model) are not that different. Nevertheless, given the observation that all considered probabilistic BW choice models basically fit empirical data equally well, longer computing times do not seem to be justified at all. Therefore, all HB-BW-CBC estimations in the Monte Carlo simulation study that will be described in the next chapter are conducted using the time-efficient MaxDiff model.

[7] All computations were based on 100.000 MCMC draws of which every 20th was stored for calculation of point estimates.
[8] All models and approaches presented in this paper were set up and programmed using the software R (see R Development Core Team 2010).

	DS 1	DS 2	DS 3
Consistent RU	**323**	**264**	**4270**
MaxDiff	**135**	**113**	**530**
Biased MaxDiff	**141**	**116**	**552**
Concordant BW	**1654**	**1363**	**--**

Table 4
Computing Times (in Minutes)

STAGE 2: SIMULATION-BASED COMPARISON OF HB-CBC, HB-BW-CBC AND THE LOUVIERE ET AL. APPROACH

Monte Carlo simulation studies are useful for comparing the relative performance of different methods without having to rely on the very specifics of usually just a few empirical data sets. The main idea is to generate "truth" – e.g. true utility parameters, choices, etc. – under a wide variety of data conditions and to assess how well different methods perform with regard to recovering the known truth. Before presenting the results of the conducted Monte Carlo simulation study, the study design and set-up will be explained.

Experimental Factors and Data Generation

The simulation design is based on the approved designs of Vriens et al. (1996) and Andrews et al. (2002a, 2002b), which also serve as a basis for numerous other simulation studies in related fields (e.g. Gensler 2003, Hillig 2006). However, taking into account insights from an extensive exploration of empirical CBC data and due to the possibility to use a high-performance computer cluster, the simulation design was both modified and extended.

The experimental factors and factor levels taken into account are displayed in *Table 5* and result in $3^5 \cdot 2 = 486$ different data conditions. Clearly, this set comprises both very challenging (e.g. 56 parameters to be estimated at the individual level using only 12 choice tasks per individual and with a very heterogeneous and small sample of individuals with large error variance) and very comfortable data situations (e.g. only 20 parameters to be estimated based on 30 choice tasks, a large and homogeneous sample and comparatively low error variances).

Factor	Abbreviation	Factor Levels
No. of choice tasks per individual	N.Ct	**12 – 18 – 30**
No. of alternatives per choice task	N.Alt	**3 – 4 – 5**
No. of parameters at the ind. level	N.Par	**20 – 35 – 56**
No. of respondents	N.Resp	**150 – 300 – 600**
Preference heterogeneity	Resp.Het	**low – high**
Error variance	Error.Var	**normal – large – individual-specific**

Table 5
Experimental Factors and Factor Levels

The process of generating the artificial data, as well as the experimental factors that are taken into account at the different process stages, are illustrated in *Figure 7*. The three main steps will be described in the following. Extensive explanations can also be found in Wirth (2010, pp. 187-198).

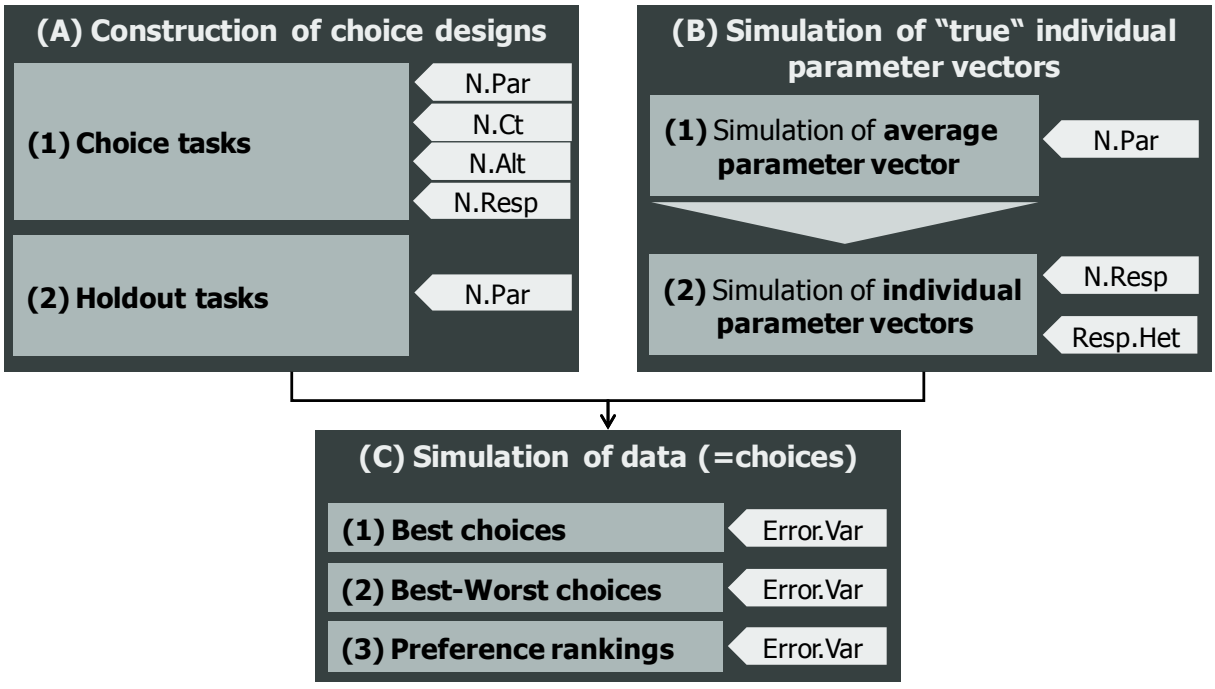


Figure 7
Data Generation

(A) Construction of Choice Designs

The choice tasks that are relevant for parameter estimation are constructed using the "Shortcut"[9] algorithm in CBC/Web (Sawtooth Software Inc. 2008, p. 15). Inputs for this algorithm are the number of choice tasks (N.Ct), the number of alternatives per choice task (N.Alt) and the number of attributes and levels. For the latter information, it is necessary to "translate" the number of parameters (N.Par) into numbers of attributes and levels. This is done as follows:

- N.Par = 20 ➔ five attributes with five levels each,
- N.Par = 35 ➔ seven attributes with six levels each,
- N.Par = 56 ➔ eight attributes with eight levels each.

Like in most of the real-world computer-administered CBC interviews, each of the artificial respondents receives an own design version. Hence, the factor N.Resp is taken into account as well, so that all in all 3^4=81 choice designs are generated.

The holdout tasks are created differently. In real-world Conjoint studies, researchers usually try to design holdout tasks in a way that does not lead to obvious and easy choices (Johnson 1997, p.1). Of course, this is also advisable in a simulation study, because obvious choices could even be predicted based on a comparatively weak estimation. As attributes and levels in simulation studies have no meaning and as preference structures are generated randomly, it is not possible to design suitable holdout tasks by applying theories about which alternatives are obvious choices. In order to make the holdout tasks difficult for the simulated respondents anyway, they are generated using the "Random" design algorithm in CBC/Web (Sawtooth Software Inc. 2008, p. 16). This way, a comparatively high degree of level overlap is ensured. Since level overlap means that the alternatives become more similar, the decisions automatically become more difficult.

Depending on the number of parameters (N.Par), ten holdout tasks with five alternatives each are generated as described above. Of course, holdout tasks are not individual-specific.

(B) Simulation of True Individual Parameter Vectors

Simulated respondents are characterized by their preference structures, i.e. by their utility parameters. In each of the 486 data scenarios, individual parameter vectors are generated as follows:

First, an N.Par-dimensional (i.e. depending on the level of the factor N.Par, a 20-, 35- or 56-dimensional) average parameter vector $\bar{\boldsymbol{\beta}}$ is generated. With the objective of reproducing typical real distributions of average partworth utilities,

- 80% of these N.Par average parameters are drawn from a Uniform distribution on the interval [2;2],
- 10% are drawn from a Uniform distribution on the interval [-5;-2] and
- 10% are drawn from a Uniform distribution on the interval [2;5].

[9] Design tests showed revealed only negligible differences in design efficiency between the "Shortcut" algorithm and the much more complex "Complete Enumeration" algorithm.

The individual utility parameters $\boldsymbol{\beta_n}$ (n=1,…,N.Resp) are generated as N.Resp random draws from a multivariate normal distribution. While the mean vector of this distribution is the average parameter vector $\bar{\boldsymbol{\beta}}$, the respective covariance matrix $\mathbf{V_\beta}$ depends on the preference heterogeneity of the simulated respondents, i.e. on the experimental factor Resp.Het. It is common in similar simulation studies to specify this covariance matrix as diagonal matrix of the form

$$\mathbf{V_\beta} = \text{var} \cdot \mathbf{I_{N.Par}}$$

with

var being a positive scalar describing the general between-subject variance and depending on the assumed preference heterogeneity and being the (N.Par×N.Par) identity matrix.

However, when analyzing empirical results of CBC studies it becomes obvious that the assumption of equal between-subject variances of all utility parameters is an unrealistic one. Instead, it is common that respondents reveal rather homogeneous preferences with regard to some features of a product while at the same time their preferences are very different regarding other product features. In order to reflect this fact more properly, the covariance matrix $\mathbf{V_\beta}$ is generated as follows:

1. A distribution of between-subject variances is simulated. This distribution depends on the experimental factor Resp.Het. In case of homogeneous preferences, the distribution consists of 10,000 random draws, each of them generated using the following R-code:

   min(rgamma(n=1,shape=0.7,scale=1.5)+runif(n=1,0.08,0.4),runif(n=1,9,11)).

   In case of heterogeneous preferences the respective distribution of between-subject variances is simulated by generating 10,000 of the following draws:

   min(rgamma(n=1,shape=0.7,scale=4.5)+runif(n=1,0.2,2),runif(n=1,13,18)).

When looking at the R-codes it becomes clear that the final distribution is a mixture of Gamma and Uniform draws. While the Gamma distributions define the general forms of the final distributions of variances10, the addition of draws from a Uniform distribution avoid the existence of too many very small variances, which would definitely occur because of the Gamma distribution having a lot of mass in the area near zero. In order to avoid variances that are too large, the resulting distribution is censored by drawing a random number from a Uniform distribution (either on the interval [9,11] or on the interval [13,18]) and then using the minimum of this draw and the sum of the Gamma and the first Uniform draw for the final distribution.

The resulting distributions of between-subject variances typically look like the ones illustrated in Figure 8.

[10] The parameters of the Gamma distributions were determined by using the distribution fitting software EasyFit (see Mathwave Technologies 2010).

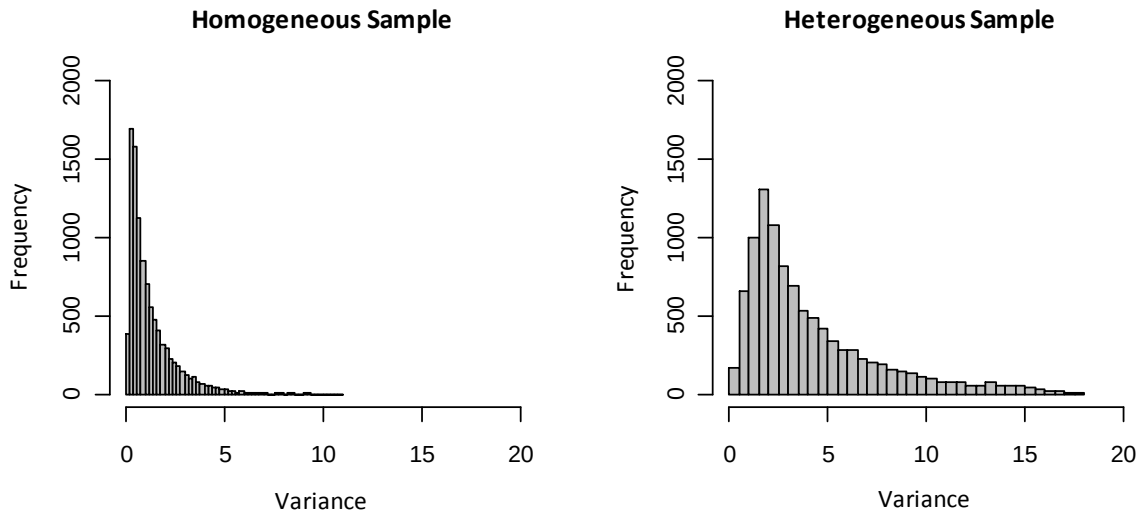


Figure 8
Distributions of Between-Subject Variances Depending on Preference Heterogeneity

2. In order to obtain the covariance matrix of interest, i.e. $\mathbf{V_\beta}$, an N.Par-dimensional vector of variances is drawn from the distribution of variances generated in step 1. With this vector being denoted as **var**, the covariance matrix is then defined as

   $$\mathbf{V_\beta} = \text{diag}(\mathbf{var}).$$

   Since the vector **var** is drawn from distributions similar to the ones in *Figure 8*, two desirable properties are ensured:

   (i) The majority of the between-subject variances is rather small, reflecting the realistic situation that respondents have similar preferences regarding many product features. However, there will also very likely be some larger values in **var**, reflecting taste heterogeneity.

   (ii) Both the probability of having larger variances in **var** and the absolute magnitude of these values increase significantly in experimental situations reflecting high preference heterogeneity (see right plot in *Figure 8*).

As already mentioned above, once $\bar{\boldsymbol{\beta}}$ and $\mathbf{V_\beta}$ are generated, the individual parameter vectors $\boldsymbol{\beta_n}$ are simply random realizations from a multivariate normal distribution with mean vector $\bar{\boldsymbol{\beta}}$ and covariance matrix $\mathbf{V_\beta}$, i.e.:

$$\boldsymbol{\beta_n} \sim \mathbf{N}(\bar{\boldsymbol{\beta}}, \mathbf{V_\beta}).$$

(C) Simulation of Data (=Choices)

Finally, data have to be simulated for all 486 scenarios and all three approaches. That is, 486 data sets with Best choices (corresponding to CBC data sets), 486 data sets with Best-Worst choices (corresponding to BW-CBC data sets), and 486 data sets with preference rankings (corresponding to Louviere et al. data sets) have to be generated.

Following Random Utility Theory (RUT), the process of simulating observations starts with calculating the individual-specific observable utility of all alternatives in all choice tasks:

$$V_{jmn} = \boldsymbol{\beta}_n' \mathbf{x}_{jmn} \qquad (j = 1, \ldots N.Alt; m = 1, \ldots N.Ct; n = 1, \ldots, N.Resp).$$

with

V_{jmn} being the observable utility of alternative j in choice task m for individual n,

$\boldsymbol{\beta}_n$ being the parameter vector of individual n and

$\mathbf{x}_{jmn}$ being the design vector for alternative j in choice task m for individual n.

Total utility U_{jmn} is then calculated by adding a Gumbel-distributed error term ε_{jmn} to V_{jmn}:

$$U_{jmn} = V_{jmn} + \varepsilon_{jmn} \qquad (j = 1, \ldots, N.Alt; m = 1, \ldots, N.Ct; n = 1, \ldots, N.Resp).$$

The variance of the Gumbel distribution is determined by the experimental factor Error.Var. If the error variance is "normal", the ε_{jmn} are drawn from

$$\varepsilon_{jmn} \sim \text{Gumbel}\left(\text{Var} = \frac{\pi^2}{6}\right),$$

and in case of a "large" error variance from

$$\varepsilon_{jmn} \sim \text{Gumbel}\left(\text{Var} = \frac{2\pi^2}{6}\right).$$

Individual-specific error variances (third level of experimental factor Error.Var) are generated by randomly drawing one value Var_n for each respondent n from a modified Gamma distribution. The associated R-code is:

$$\mathrm{Var}_n = \mathrm{rgamma}(n = 1, \mathrm{shape} = 2, \mathrm{scale} = 0.6) + \mathrm{runif}(n = 1, 0.9, 1.3).$$

The resulting distributions of individual error variances typically look like the one illustrated in *Figure 9.* While the expected value of the individual error variances generated this way is 2.3, the graph shows that the individual realizations are skew-distributed, may well differ a lot and contain both pretty small and comparatively large values. All these properties theoretically pose challenges for HB choice algorithms.

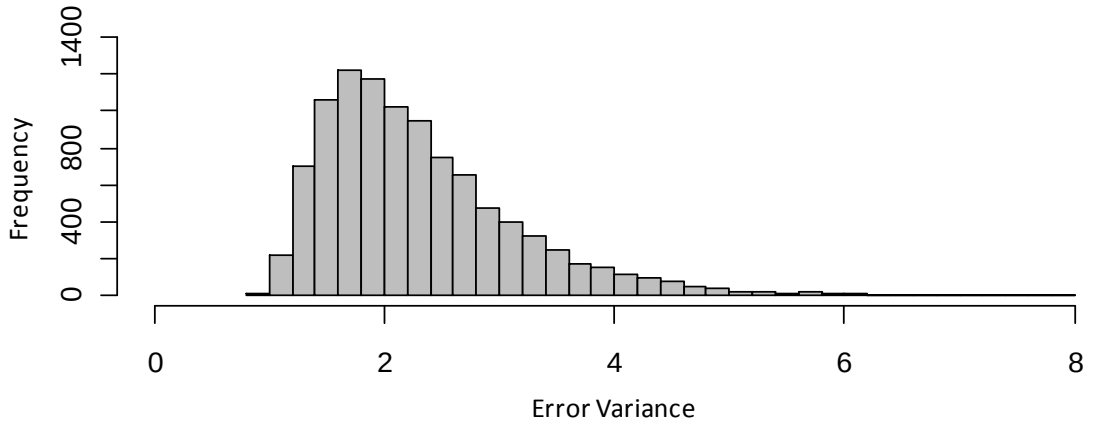


Figure 9
Distribution of Individual Error Variances

Based on the calculated total utilities U_{jmn}, the observations in the choice tasks are simulated as follows:

- ***HB-CBC (Best choices only)***: In each choice task and for each respondent, the alternative with highest total utility is the chosen one.
- ***HB-BW-CBC (Best-Worst choices)***: The pair of alternatives with the maximum difference in total utility is the chosen Best-Worst pair in each choice task and for each respondent.[11]
- ***Louviere et al. approach (preference ranking)***: In each choice task and for each respondent, the alternatives are ranked with respect to their total utility.[12]

As in the end researchers are usually interested in predicting purchases and not Best-Worst choices or preference rankings, only Best choices are simulated for the ten holdout tasks.

[11] The chosen procedure for generating the Best-Worst data does not perfectly agree with the theoretical MaxDiff model. According to the correct model (see e.g. Marley, Louviere 2005, p. 471), the Gumbel-distributed error had to be added to differences in observed utilities, not to the alternative-specific observed utilities themselves. However, if this correct MaxDiff model had been implemented, the random proportion of each alternative's total utility would be lower in the data that is used by HB-BW-CBC than in the data used by the other two approaches. Therefore, the slight deviation from the assumed data-generating process in MaxDiff was accepted in order not to favor HB-BW-CBC in the simulation-based comparison.

[12] There is no explicit assumption regarding the process that generates the rank orders in Louviere et al. (2008). Therefore, the simulation of the data is based on the plausible assumption that rank orders are generated according to the total utilities of the ranked alternatives.

MEASURES OF PERFORMANCE

The 1,458 data sets generated as described above are the basis for the comparison of HB-CBC, HB-BW-CBC and the Louviere et al. approach. The main idea is simple: First, each approach is used to estimate individual utility parameters based on "its" 486 data sets. The results of the estimations are then compared with regard to the popular and widely used performance measures that are displayed in *Table 6*.

The average[13] correlation between the true and the estimated individual parameters and the correlation between true and estimated attribute importances[14] are the used measures of parameter recovery. Clearly, the higher the correlation coefficients, the better the match between estimated and true parameters.

	Parameter Recovery	**(Internal) Predictive Validity**
Individual Level	Average Pearson Correlation between true and estimated individual utility parameters	First Choice Hit Rate in holdout tasks
Aggregate Level	Pearson Correlation between true and estimated attribute importances	Mean Absolute Error (MAE) between true and estimated preference shares in holdout tasks

Table 6
Performance Measures Used For Simulation-Based Comparison

The First Choice Hit Rate, which is the proportion of correctly predicted Best choices in the holdout tasks, is used to assess predictive validity at the individual level. At the aggregate level, predictive validity is measured by the Mean Absolute Error (MAE) between the true and the estimated preference shares in the holdout tasks. A higher Hit Rate reflects a better prediction of individual choices. A lower MAE corresponds to a better prediction of aggregate choice shares - no matter if individual choices are predicted correctly or not. While both measures are very popular and widely used for assessing predictive accuracy of Conjoint approaches, it is important to keep in mind that strictly speaking both the Hit Rate and the MAE as used here are measures of *internal* predictive validity only. This is because they reflect how well an approach predicts choices made by the respondents in the choice experiment rather than how well they predict real choices or preference shares calculated based on another sample.

RESULTS

Due to the huge amount of information – all in all there are 1,458 estimations evaluated with regard to four different performance measures – a systematic approach is required for the simulation-based comparison.

First, four ANOVAs are conducted, with the respective performance measure as dependent variable and the experimental factors, a variable indicating the approach (HB-CBC, HB-BW-

[13] Before averaging the N.Resp correlations, Fisher's Z-transformation is applied (see Silver, Dunlap 1987).
[14] As suggested by e.g. Orme (2006, p. 80), the overall importance of an attribute is calculated as the average of the respective individual attribute importances.

CBC or Louviere et al.) and the interactions between the approach and the experimental factors as independent variables. The F- and p-values of the main effects reveal if the experimental factors and the approach have an influence on the respective performance measure. The F- and p-values of the interaction effects show if the influence of the experimental factors on the performance measures depends on the approach. More detailed insights can be gained by looking at the means of the performance measures for each method and under each level of the experimental factors. Since it turns out that the ANOVAs do not yield additional valuable information, the following discussion of the results focuses only on these means. Nevertheless, concentrated information about the ANOVAs can be found in *Table 13* and *Table 14* in the appendix.

Parameter Recovery

The average correlations between the true and the estimated individual parameters by model type (HB-CBC, HB-BW-CBC, Louviere et al.) and experimental condition are summarized in *Table 7*.

	HB-CBC (I)	HB-BW-CBC (II)	Louviere (III)
N.Ct			
12 (a)	0.843	0.884	0.725
18 (b)	0.870 ^{a**}	0.909 ^{a**}	0.799 ^{a**}
30 (c)	0.903 ^{a**,b**}	0.937 ^{a**,b**}	0.873 ^{a**,b**}
N.Alt			
3 (a)	0.862	0.895	0.725
4 (b)	0.875 ^{a**}	0.915 ^{a**}	0.810 ^{a**}
5 (c)	0.884 ^{a**,b**}	0.926 ^{a**,b**}	0.866 ^{a**,b**}
N.Par			
20 (a)	0.908 ^{b**,c**}	0.941 ^{b**,c**}	0.876 ^{b**,c**}
35 (b)	0.871 ^{c**}	0.910 ^{c**}	0.799 ^{c**}
56 (c)	0.834	0.876	0.719
N.Resp			
150 (a)	0.865	0.908	0.808
300 (b)	0.874 ^{a**}	0.913 ^{a**}	0.808
600 (c)	0.883 ^{a**,b**}	0.918 ^{a**,b**}	0.806
Resp.Het			
Low (a)	0.904 ^{b**}	0.930 ^{b**}	0.797
High (b)	0.836	0.892	0.818 ^{a*}
Error.Var			
Normal (a)	0.878 ^{b**}	0.916 $^{b**,[c*]}$	0.821 ^{b**,c**}
Large (b)	0.870	0.908	0.790
Ind.-spec. (c)	0.875 ^{b*}	0.914 ^{b**}	0.810 ^{b**}
Overall	**0.874** ^{III**}	**0.913** ^{I**,III**}	**0.808**

Table 7
Means of Average Correlations Between True and Estimated Individual Parameters by Model Type and Experimental Condition

Before going into details, some general remarks about the interpretation of the mean table (and all the similar tables that are to be shown in the following) are required:

Factor level means are tested for significant pairwise differences within each method. The overall performance of the three approaches regarding the performance measure can be evaluated by looking at the overall means that are printed in the last line of the table. These means are tested for significant differences, too. If pairwise differences are found to be significant, superscripts are added to the superior mean as follows:

- Small letter(s) (respectively, roman number(s)) denote the significantly inferior factor level(s) (respectively, the significantly inferior approach(es)).
- One star (*) denotes differences that are significant at the 0.05 level,
- two stars (**) denote differences that are significant at the 0.01 level.
- Deviations between significances found based on uncorrected t-tests and t-tests using the Bonferroni-adjustment are marked by square brackets. For example, $a^{*[*]}$ would mean that the difference between the mean looked at and the mean of the first factor level is significant at the 0.01 level based on an unadjusted t-test, but only at the 0.05 level when using the Bonferroni adjustment.

When looking at the values in Table 7, the following conclusions can be drawn regarding the recovery of individual utility parameters:[15]

- Clearly and as expected, all three approaches benefit from better data conditions. The more choice tasks and alternatives per choice task there are and the fewer parameters have to be estimated, the better the estimated parameters match the true ones.
- It is remarkable how well both HB approaches perform even in very sparse data conditions. In these situations, they are clearly superior to the Louviere et al. approach.
- Nevertheless, the Louviere et al. approach benefits a lot from an improvement of data conditions: While it recovers individual parameters much worse than the HB approaches in situations that are characterized by e.g. very few choice tasks or very many parameters, its performance almost equals the performance of HB-CBC when data conditions become comfortable.
- It can also be seen that the "borrowing strength" property (see Bradlow et al. 2005, p. 33) of HB models leads to the two HB approaches benefitting slightly from a larger sample. However, it should be noted that sample size seems to have a much lesser influence on parameter recovery than choice design parameters, like e.g. the number of choice tasks.
- The more heterogeneous the sample and larger the error variance is, the worse is the HB approaches' performance regarding the recovery of true individual utility parameters. Nevertheless, individual-specific variances do not seem to harm the ability of the HB approaches to correctly recover individual utility parameters. This may be surprising to some researchers, as the standard HB models implemented here do not explicitly account for different individual error variances.

[15] Only the most important findings are summarized. An extensive discussion of all results can be found in Wirth (2010, pp. 203-232).

- All in all, HB-BW-CBC significantly outperforms HB-CBC with regard to the recovery of true individual parameters. Although the Louviere et al. approach is doing pretty well in comfortable data situations, its problems in challenging data conditions lead to the approach being on average clearly inferior to both HB-based alternatives.

The recovery of aggregate preference structures is assessed based on correlations between the true and the estimated attribute importances. *Table 8* displays the respective means.

	HB-CBC (I)	HB-BW-CBC (II)	Louviere (III)
N.Ct			
12 (a)	0.980	0.992	0.988
18 (b)	0.986 ^{a**}	0.995 ^{a**}	0.995 ^{a**}
30 (c)	0.994 ^{a**,b**}	0.998 ^{a**,b**}	0.997 ^{a**,b**}
N.Alt			
3 (a)	0.986	0.994	0.989
4 (b)	0.988	0.996 ^{a**}	0.995 ^{a**}
5 (c)	0.990 ^{a**}	0.997 $^{a**,b*[*]}$	0.997 ^{a**,b**}
N.Par			
20 (a)	0.995 ^{b**,c**}	0.999 ^{b**,c**}	0.998 ^{b**,c**}
35 (b)	0.987 ^{c**}	0.995 ^{c**}	0.994 ^{c**}
56 (c)	0.974	0.989	0.985
N.Resp			
150 (a)	0.984	0.993	0.992
300 (b)	0.988 $^{a*[*]}$	0.996 ^{a**}	0.994 ^{a**}
600 (c)	0.991 $^{a**,b*[*]}$	0.997 ^{a**,b**}	0.997 ^{a**,b**}
Resp.Het			
Low (a)	0.993 ^{b**}	0.997 ^{b**}	0.995
High (b)	0.979	0.993	0.994
Error.Var			
Normal (a)	0.988	0.996	0.995
Large (b)	0.988	0.996	0.994
Ind.-spec. (c)	0.988	0.996	0.995
Overall	**0.988**	**0.996** ^{I**,III**}	**0.994** ^{I**}

Table 8
Means of Correlations Between True and Estimated Attribute Importances by Model Type and Experimental Condition

It can be seen that, irrespective of the data situation, the recovery of attribute importances does not pose a problem for any of the three approaches. Hence, there is no need to go too much into details here. HB-BW-CBC has the best overall performance again, but this time the Louviere et al. approach is the second-best approach with a significantly better recovery of attribute importances than standard HB-CBC. Finally, it is worth noting that individual-specific error variances do not negatively influence the performance of both HB approaches regarding this performance measure, too.

Predictive Validity

The means of the First Choice Hit Rate are summarized in *Table 8*. The most important findings are similar to the ones regarding individual parameter recovery:

- Once more and not surprisingly, all approaches benefit from better data conditions, but the Louviere et al. approach benefits by far the most.
- The HB approaches yield very satisfactory results and are clearly superior to the Louviere et al. approach in challenging data situations, but the difference becomes smaller when data conditions become more comfortable.
- While a larger sample size has a slightly positive effect on individual parameter recovery of the HB approaches (see above), this effect seems to be too small to translate into better First Choice Hit Rates. Hence, it becomes even more obvious that for HB approaches sample size is less a driver of estimation performance than properties of the choice design, such as the number of choice tasks.
- The more heterogeneous the sample and the larger the error variance is, the worse is the prediction of individual choice behavior when using HB approaches.
- Like the other measures presented so far, the individual-level predictive validity of the HB approaches is not negatively affected by individual-specific error variances.
- Overall, both HB-based alternatives are superior to the Louviere et al. approach with respect to the First Choice Hit Rate. Among the HB approaches, HB-BW-CBC is consistently superior to HB-CBC.

	HB-CBC (I)	HB-BW-CBC (II)	Louviere (III)
N.Ct			
12 (a)	62.15%	65.55%	54.39%
18 (b)	64.66% [a**]	68.19% [a**]	59.40% [a**]
30 (c)	67.58% [a**,b**]	71.24% [a**,b**]	65.30% [a**,b**]
N.Alt			
3 (a)	63.33%	66.33%	53.83%
4 (b)	64.86% [a**]	68.51% [a**]	60.13% [a**]
5 (c)	66.20% [a**,b**]	70.15% [a**,b**]	65.14% [a**,b**]
N.Par			
20 (a)	66.70% [c**]	69.59% [c**]	63.72% [b**,c**]
35 (b)	65.74% [c**]	69.49% [c**]	60.85% [c**]
56 (c)	61.95%	65.90%	54.52%
N.Resp			
150 (a)	64.59%	68.18%	60.12% [c*]
300 (b)	64.60%	68.26%	59.63%
600 (c)	65.20%	68.55%	59.34%
Resp.Het			
Low (a)	66.69% [b**]	69.18% [b**]	58.12%
High (b)	62.91%	67.48%	61.28% [a**]
Error.Var			
Normal (a)	66.50% [b**]	70.27% [b**,c*[*]]	61.76% [b**,c**]
Large (b)	62.36%	65.78%	57.01%
Ind.-spec. (c)	65.53% [b**]	68.93% [b**]	60.32% [b**]
Overall	**64.80%** [III**]	**68.33%** [I**,III**]	**59.70%**

Table 9
Means of First Choice Hit Rate by Model Type and Experimental Condition

The final performance measure of interest is the Mean Absolute Error (MAE) between the true and the estimated preference shares in the holdout tasks. The preference shares were calculated using both the First Choice rule and the Logit rule. As the results and findings are very similar, the following explanations focus on the Logit rule variant. The means of the MAE are displayed in *Table 10*.

	HB-CBC (I)	HB-BW-CBC (II)	Louviere (III)
N.Ct			
12 (a)	5,60%	4,38%	4,83%
18 (b)	4,71% [a**]	3,58% [a**]	3,54% [a**]
30 (c)	3,68% [a**,b**]	2,73% [a**,b**]	2,35% [a**,b**]
N.Alt			
3 (a)	5,06%	3,90%	4,87%
4 (b)	4,63% [a**]	3,47% [a**]	3,36% [a**]
5 (c)	4,31% [a**,b**]	3,32% [a**,[b*]]	2,49% [a**,b**]
N.Par			
20 (a)	3,61% [b**,c**]	2,76% [b**,c**]	2,04% [b**,c**]
35 (b)	4,72% [c**]	3,50% [c**]	3,38% [c**]
56 (c)	5,67%	4,43%	5,31%
N.Resp			
150 (a)	4,90%	3,74%	3,84%
300 (b)	4,67% [a*[*]]	3,55% [a**]	3,49% [a**]
600 (c)	4,43% [a**,b*[*]]	3,40% [a**,[b*]]	3,38% [a**]
Resp.Het			
Low (a)	5,07%	4,06%	4,01%
High (b)	4,26% [a**]	3,07% [a**]	3,13% [a**]
Error.Var			
Normal (a)	4,57% [[b*],[c*]]	3,55%	3,60%
Large (b)	4,71%	3,58%	3,55%
Ind.-spec. (c)	4,72%	3,56%	3,57%
Overall	**4,67%**	**3,56%** [I**]	**3,57%** [I**]

Table 10
Means of MAE (Basis: Logit Rule) Between True and Estimated Preference Shares in Holdout Task by Model Type and Experimental Condition

The results with respect to the MAE differ somewhat from the findings regarding the other performance measures and can be summarized as follows:

- While better data conditions generally result in lower MAEs, the Louviere et al. approach does not only profit a lot more from an improvement of the data situation than the other alternatives in terms of the MAE – this time it also yields very satisfactory results in challenging situations. Under very good data conditions, the Louviere et al. approach even yields the lowest MAE of all approaches considered.
- As the MAE is calculated based on the estimated preference share and since the standard error of this estimate decreases when sample size is increased, a larger sample generally leads to smaller MAEs.
- The MAE is the first considered measure that profits from more heterogeneous preference structures when parameters are estimated using HB approaches. A detailed analysis of the results has revealed that one potential reason for this phenomenon might be Bayesian shrinkage, which is particularly strong in these situations. While shrinkage of individual parameters towards the population mean is obviously detrimental when

looking at individual performance measures, such as the Hit Rate, it could be helpful when aggregate measures are of interest.

- Finally, it can be observed that the error variance has almost no effect on the MAE. This can be explained by the MAE being based on the average of individual choice probabilities: While larger error variances result in more statistical noise and thus in larger errors when estimating individual choice probabilities, these errors are non-systematic so that they cancel out when averaging the individual estimates.
- Once more, the performance of the HB approaches is not negatively affected by the presence of individual-specific error variances.
- Overall, the Louviere et al. approach reaches the performance level of HB-BW-CBC. Both approaches are significantly better in predicting preference shares than the standard HB-CBC approach.

Summary of the Results of the Monte Carlo Simulation Study

Based on the findings of the Monte Carlo simulation study, the following characteristics of the three approaches can be derived:

In spite of its simplicity, the Louviere et al.-approach leads on average to a good recovery of attribute importances and a good prediction of preference shares (=aggregate performance measures). While its unique characteristic – the purely individual estimation – seems to be detrimental when data conditions are challenging, the approach benefits the most from an improvement of the data situation, like e.g. a greater number of choice tasks or fewer parameters to be estimated. Irrespective of the data conditions, both HB approaches are superior in terms of performance measures at the individual level (Hit Rate, recovery of individual utility parameters).

The standard HB-CBC approach is characterized by a good recovery of individual parameters and a good prediction of individual choices. Other than in the study of Andrews et al. (2002a) the approach yields remarkable results even in very challenging data conditions. However, regarding aggregate performance measures it is inferior to the two alternative approaches.

The additional preference information extracted from the Worst choice leads to the HB-BW-CBC approach being the best of the considered alternatives in terms of the recovery of individual parameters and aggregate attribute importances as well as in terms of the prediction of individual choices. Furthermore, HB-BW-CBC is characterized by the best prediction of aggregate preference shares whenever data conditions are very challenging.

From a methodological point of view it remarkable that the hypothesized systematically negative influence of individual-specific, skew-distributed error variances on the performance of the two HB-approaches could not be confirmed at all in this simulation study.

All in all, one can conclude from the simulation study that the Louviere et al.-approach is an alternative worth considering when data conditions are not too challenging and/or the main interest of the study is the prediction of aggregate preference shares. Its conceptual and methodological simplicity makes it particularly attractive for practitioners who regard the complex HB algorithms as black boxes. However, the results suggest that whenever the data

situation is not good, HB approaches should be preferred. This conclusion is supported by the fact that the often hypothesized dramatic sensitivity of the performance of HB choice models to individual-specific error variances could not be confirmed. The results of the simulation study also suggest that the additional information that is extracted from Best-Worst choice tasks may indeed lead to a superior estimation of individual parameters even in very sparse data conditions. For example, irrespective of the performance measure you look at, HB-BW-CBC performs better in data situations with only 12 choice tasks than HB-CBC does in data situations characterized by 18 choice tasks.

In order to see whether the superiority of HB-BW-CBC to HB-CBC also can be confirmed based on real data, an empirical comparison of these two approaches was conducted. The results are presented in the following section.

Some Empirical Insights on the Relative Performance of HB-CBC and HB-BW-CBC

In 2009, GfK Marketing Sciences conducted four Best-Worst CBC studies. The characteristics of the data sets are displayed in *Table 11*. It can be seen that the underlying choice experiments were comparatively simple.

Data Set	No. of Choice Tasks	No. of Attributes (Parameters)	No. of Respondents
Data Set 1	**6**	**4 (10)**	**206**
Data Set 2	**5**	**4 (10)**	**206**
Data Set 3	**4**	**3 (4)**	**206**
Data Set 4	**5**	**3 (8)**	**206**

Table 11
Characteristics of the Data Sets Used for the Empirical Comparison

The study was originally not designed as a side-by-side test of HB-CBC and HB-BW-CBC. Hence, it is not a split sample survey and there is no pre-defined holdout task. In order to be able to assess the additional value of asking for best ***and worst*** instead of only asking for best anyway, the following was done:

1. The last choice task in each data set was used as a holdout task.
2. Two sets of parameters were estimated:
    a. one using HB-CBC, i.e. considering only Best choices and
    b. one using HB-BW-CBC, i.e. considering Best-Worst choices.
3. First Choice Hit Rates were calculated based on these two parameter sets.

These Hit Rates are displayed in *Table 12*. It can be observed that HB-BW-CBC leads to a better prediction of individual choice behavior in each of the four data sets. Taking into account that the data sets describe rather straightforward choice experiments and that HB-CBC already

achieves very satisfying Hit Rates, the additional 4-5 percentage points in Hit Rate achieved by HB-BW-CBC in Data Sets 1, 3 and 4 are impressive.

Data Set	HB-BW-CBC	HB -CBC
Data Set 1	**69.9 %**	**65.5 %**
Data Set 2	**63.6 %**	**62.1 %**
Data Set 3	**77.7 %**	**73.3 %**
Data Set 4	**76.2 %**	**71.3 %**

Table 12
First Choice Hit Rates

Of course, these findings are no final proof for the empirical superiority of HB-BW-CBC - the data sets were too similar to allow a generalization. Nevertheless, the remarkable fact that the empirical results confirm the respective conclusions of the Monte Carlo simulation study so clearly is a strong indication that asking for the best and the worst alternative in each choice task and adapting the modeling accordingly may likely improve parameter estimation and predictive power of choice models.

Summary and Outlook

The findings presented in this paper suggest that the purely individual Louviere et al. approach is a simple alternative to standard CBC that is well worth considering in some situations, but that there are also good reasons for the popularity of the HB-CBC approach: It consistently yields good results, even under sparse data conditions and even when individual error variances are not constant across the sample.

Nevertheless, Best-Worst CBC outperforms standard CBC with regard to all considered performance measures in the simulation study. This superiority can also be confirmed based on four empirical data sets. Therefore, it seems to be justified to conclude that Best-Worst CBC is an alternative to standard CBC that practitioners should consider, in particular when data conditions are challenging. Empirical evidence presented in this paper shows that the relatively simple MaxDiff model does a good job in fitting Best-Worst data, so barriers for switching from CBC to BW-CBC are not too high.

The flexibility of Marley & Louviere's (2005) probabilistic Best-Worst choice models provides starting points for future research, too. For instance, insights from further empirical research could be used to specify Best-Worst choice models that match the true data generating process even better than the MaxDiff model and the other models that were taken into account in the presented research project. In addition, further empirical comparison studies of HB-CBC and HB-BW-CBC, preferably based on challenging data sets that differ as much as possible from the four data sets used in the empirical comparison presented above, will help to assess the approaches' strengths and weaknesses in even more detail. Finally, whenever possible, researchers should try to validate the approaches based on measures of internal validity, such as

the prediction of choice behavior in holdout tasks, ***and*** measures of external validity, such as the prediction of actual market shares.

REFERENCES

Andrews, Rick L.; Ainslie, Andrew; Currim, Imran S. (2002a): An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity. In: Journal of Marketing Research 39(4), 479–487.

Andrews, Rick L.; Ansari, Asim; Currim, Imran S. (2002b): Hierarchical Bayes versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction and Partworth Recovery. In: Journal of Marketing Research 39(1), 87–98.

Bradlow, Eric T.; Lenk, Peter J.; Allenby, Greg M.; Rossi, Peter E. (2005): When BDT in Marketing Meant Bayesian Decision Theory: The Influence of Paul Green's Research. In: Wind, Yoram; Green, Paul E (eds.): Marketing Research and Modeling: Progress and Prospects. A Tribute to Paul E. Green. New York, NY: Springer (International Series in Quantitative Marketing, 14), 17–39.

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

Chrzan, Keith; Fellerman, Ritha (1997): A Comparison of Full- and Partial-Profile with Best-Worst Conjoint Analysis. In: Sawtooth Software Inc. (ed.): 2007 Sawtooth Software Conference Proceedings, 59-67.

Congdon, Peter (2003): Bayesian Statistical Modelling. Chichester: Wiley (Wiley Series in Probability and Statistics).

Diener, Christopher G.; Orme, Bryan K.; Yardley, Dan (2006): Dual Response "None" Approaches: Theory And Practice. In: Sawtooth Software Inc. (ed.): 2006 Sawtooth Software Conference Proceedings, 157–167.

Gensler, Sonja (2003): Heterogenität in der Präferenzanalyse. Ein Vergleich von hierarchischen Bayes-Modellen und Finite-Mixture-Modellen. Wiesbaden: Dt. Univ.-Verl. (Beiträge zur betriebswirtschaftlichen Forschung, 107).

Hillig, Thomas (2006): Verfahrensvarianten der Conjoint-Analyse zur Prognose von Kaufentscheidungen. Eine Monte-Carlo-Simulation. Wiesbaden: Dt. Univ.-Verl. (Gabler Edition Wissenschaft).

Johnson, Richard M. (1997): Including Holdout Choice Tasks in Conjoint Studies. Sawtooth Software Inc. Sequim, WA. Sawtooth Software Research Paper Series. Online: http://www.sawtoothsoftware.com/download/techpap/inclhold.pdf.

Louviere, Jordan J.; Eagle, Thomas C. (2006): Confound It! That Pesky Little Scale Constant Messes Up Our Convenient Assumptions. In: Sawtooth Software Inc. (ed.): 2006 Sawtooth Software Conference Proceedings, 211–228.

Louviere, Jordan J.; Street, Deborah J.; Ainslie, Andrew; Deshazo, J. R.; Cameron, Trudy; Hensher, David A. et al. (2002): Dissecting the Random Component of Utility. In: Marketing Letters 13(3), 177–193.

Louviere, Jordan J.; Street, Deborah J.; Burgess, Leonie; Wasi, Nada; Islam, Towhidul; Marley, Anthony A. J. (2008): Modeling the Choices of Individual Decision-Makers by Combining Efficient Choice Experiment Designs with Extra Preference Information. In: Journal of Choice Modeling 1(1), 128–163.

Marley, Anthony A. J. (1968): Some Probabilistic Models of Simple Choice and Ranking. In: Journal of Mathematical Psychology 5, 311–332.

Marley, Anthony A. J. (2010): The best–worst method for the study of preferences. In: Frensch, Peter A.; Schwarzer, Ralf (eds.): Cognition and Neuropsychology: International Perspectives on Psychological Science, Volume 1. London: Psychology Press, 147-157.

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.

Mathwave Technologies (2010): EasyFit. Distribution Fitting Software, Version 5.0. Online: http://www.mathwave.com/easyfit-distribution-fitting.html.

Moore, William L. (2004): A Cross-Validity Comparison of Rating-Based and Choice-Based Conjoint Analysis Models. In: International Journal of Research in Marketing 21(3), 299–312.

Moore, William L.; Gray-Lee, Jason; Louviere, Jordan J. (1998): A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation. In: Marketing Letters 9(2), 195–207.

Newton, Michael A.; Raftery, Adrian E. (1994): Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. In: Journal of the Royal Statistical Society (B) 56(1), 3–48.

Orme, Bryan K. (2006): Getting Started with Conjoint Analysis. Strategies for Product Design and Pricing Research. Madison, Wis.: Research Publ.

Pinnell, Jon (2004): Comment on Huber: Practical Suggestions for CBC Studies. In: Sawtooth Software Inc. (ed.): 2004 Sawtooth Software Conference Proceedings, 43–52.

Poynter, Ray (2006): The Power of Conjoint Analysis and Choice Modelling in Online Surveys. Market Research Society, Annual Conference, Barbican, London. Online: http://www.virtualsurveys.com/node/231.

R Development Core Team (2010): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Online: http://www.R-project.org/.

Rossi, Peter E.; Allenby, Greg M.; McCulloch, Robert E. (2005): Bayesian Statistics and Marketing. Hoboken, NJ: Wiley (Wiley Series in Probability and Statistics).

Sawtooth Software Inc. (2008): CBC v6.0 Technical Paper. Sequim, WA. Sawtooth Software Technical Paper Series. Online: http://www.sawtoothsoftware.com/download/techpap/cbctech.pdf.

Sawtooth Software Inc. (2010): Report on Conjoint Usage. Sawtooth Solutions Newsletter, Spring 2010. Online: http://www.sawtoothsoftware.com/education/ss/ss31.shtml#ss31report.

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.

Swait, Joffre; Louviere, Jordan J.; Anderson, Don (1995) Best/Worst Conjoint: A New Preference Elicitation Method to Simultaneously Identify Overall Attribute Importance and Attribute Level Partworths. Working Paper.

Teichert, Thorsten (2001): Nutzenermittlung in wahlbasierter Conjoint-Analyse: Ein Vergleich zwischen Latent-Class- und hierarchischem Bayes-Verfahren. In: Zeitschrift für betriebswirtschaftliche Forschung 53, 798–822.

Vriens, Marco; Wedel, Michel; Wilms, Tom (1996): Metric Conjoint Segmentation Methods: A Monte Carlo Comparison. In: Journal of Marketing Research 33(1), 73–85.

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

APPENDIX

RESULTS OF THE ANOVAS CONDUCTED IN THE CONTEXT OF THE SIMULATION STUDY:

	F-Ratio (Corr. Util.)	p	F-Ratio (Corr. Imp.)	p
N.Ct	**3081.70**	**0.0000**	**246.25**	**0.0000**
N.Alt	**1385.97**	**0.0000**	**85.43**	**0.0000**
N.Par	**4009.84**	**0.0000**	**472.95**	**0.0000**
N.Resp	**49.81**	**0.0000**	**85.36**	**0.0000**
Resp.Het	**1872.37**	**0.0000**	**209.23**	**0.0000**
Error.Var	**91.34**	**0.0000**	1.61	0.2006
Approach	**4896.29**	**0.0000**	**158.89**	**0.0000**
Approach×N.Ct	**71.02**	**0.0000**	**3.15**	**0.0137**
Approach×N.Alt	**222.07**	**0.0000**	**8.55**	**0.0000**
Approach×N.Par	**44.06**	**0.0000**	**2.40**	**0.0485**
Approach×N.Resp	**16.18**	**0.0000**	2.29	0.0578
Approach×Resp.Het	**933.22**	**0.0000**	**45.35**	**0.0000**
Approach×Error.Var	**8.10**	**0.0000**	0.54	0.7061
R^2 (adj. R^2)	**0.958 (0.957)**		**0.635 (0.626)**	

Table 13
F-Tests of Main and Interaction Effects on Parameter Recovery Measures

	F-Ratio (Hit Rate)	p	F-Ratio (MAE)	p
N.Ct	**376.91**	**0.0000**	**810.59**	**0.0000**
N.Alt	**252.15**	**0.0000**	**315.73**	**0.0000**
N.Par	**265.82**	**0.0000**	**1094.47**	**0.0000**
N.Resp	0.29	0.7496	**37.17**	**0.0000**
Resp.Het	**12.52**	**0.0004**	**482.22**	**0.0000**
Error.Var	**147.91**	**0.0000**	0.54	0.5821
Approach	**525.43**	**0.0000**	**321.63**	**0.0000**
Approach×N.Ct	**22.33**	**0.0000**	**12.27**	**0.0000**
Approach×N.Alt	**49.88**	**0.0000**	**68.15**	**0.0000**
Approach×N.Par	**20.08**	**0.0000**	**49.21**	**0.0000**
Approach×N.Resp	1.34	0.2546	0.74	0.5624
Approach×Resp.Het	**88.65**	**0.0000**	1.55	0.2121
Approach×Error.Var	0.25	0.9072	0.81	0.5209
R^2 (adj. R^2)	**0.722 (0.716)**		**0.813 (0.808)**	

Table 14
F-Tests of Main and Interaction Effects on Predictive Validity Measures

COMMENT ON MARSHALL ET AL. AND WIRTH

BRYAN ORME
SAWTOOTH SOFTWARE, INC.

MOTIVATION

The primary aim of these two presentations (Marshall et al. and Wirth) was to test the bold assertions that Jordan Louviere made at the 2009 Sawtooth Software Conference. Louviere argued that traditional CBC models (where respondents pick the best alternative from sets) modeled with HB estimation were inferior and biased. He cited evidence from split-sample studies (though the HB estimation was programmed by his team, rather than provided by commercial CBC/HB software). He proposed a new approach (Bottom Up) that collected more than just first choices and used purely individual-level estimation. He put us on notice in 2009, announcing that our common Top Down methods were "soooo WRONG," and that Bottom Up methods were like an asteroid strike that would lead to species extinction.

CREDIT WHERE CREDIT IS DUE

Jordan should be given credit for the many contributions to the field, especially his influential paper in 1983 that demonstrated to the marketing community the benefits and mechanics of discrete choice experiments. Jordan's MaxDiff scaling was also a very useful invention. For these contributions and others, Louviere was awarded the 2010 Parlin Award. On a personal note, Jordan has been very helpful and patient with me as he has answered detailed emails regarding MaxDiff scaling and other related issues.

Jordan has correctly argued that individual respondents shouldn't be directly compared (on the utilities) without somehow accounting for scale differences. Sawtooth Software's founder, Rich Johnson, recognized this issue as early as the 1970s. Since the 1980s, Sawtooth Software's market simulators have summarized respondent utilities for reporting purposes after applying a normalization procedure. For each respondent, a normalizing constant is selected such that the sums of utility ranges across attributes are equal for each respondent. In our most recent simulators, this normalization procedure is called *zero-centered diffs*.

Sawtooth Software advocates using zero-centered diffs in tabulations when comparing groups and also in subsequent cluster analyses to find groups of similar respondents. But, raw utilities are used in the market simulator to project respondent choices.

WAS JORDAN RIGHT IN HIS 2009 PRESENTATION?

Why was an entire session of this conference dedicated to the subject of Bottom-Up vs. Top-Down methods? Jordan's 2009 presentation criticizing the traditional CBC approach with HB (Top-Down) motivated the audience to award him the Best Paper presentation. So, he apparently captured your attention!

In Jordan's 2009 presentation, he included the following in his PowerPoint slides:

- The world that you knew has changed & will never again be the same.
- Current choice models are WRONG!
- They are soooo WRONG, it's hard to know why so many folks keep working on them.
- All published empirical results are WRONG & should be in the rubbish bin of failed science.
- Stop using these models NOW!

Jordan described his Bottom-Up approach as a game-changing asteroid event, akin to the massive global strike 60 million years ago that is argued to have led to the extinction of the dinosaurs. These were indeed bold assertions, that if correct, would have meant that those using traditional CBC and HB estimation were harming their clients and risked extinction.

Thanks to the Herculean efforts of Ralph Wirth, Joe Curry, Don Marshall, Siu-Shing Chan, Rich Johnson, Jordan Louviere, Bart Frischknecht, and John Rose, we now have substantial evidence that Jordan was not right.

Wirth's synthetic data studies suggest that if you want to use Jordan's questionnaire approach for CBC, there is no advantage to using the purely individual-level estimation over HB. Even when Wirth varied the error variance across respondents, he found no troubles for HB. The claims that HB estimation is biased and misleading seem unfounded. The recovery of known utility parameters was solid and unbiased for HB. In Jordan's rebuttal at the conference, he dismissed Wirth's findings, and said that the latest version of his BU estimation routine (as used by Marshall et al.) is superior to the previous version Wirth employed. That may be true, but it does not change Wirth's findings with respect to HB recovery of true parameters.

Marshall *et al.*'s two studies (pizza and camera) suggest that for real respondents, Bottom-Up doesn't do any better than traditional Top-Down CBC (for the camera data set, it was generally worse). But, Bottom-Up…

- Requires more data
- Takes much longer respondent effort
- More respondents drop out in BU
- More respondents are dissatisfied with the BU survey
- No commercial or open source software is available for BU

And, I'd like to add that Bottom-Up methods would especially be at a disadvantage when many attributes and levels are involved, attribute interactions are significant, and the survey is already relatively long.

Jordan's Bottom-Up approach has two major differences from Sawtooth Software's standard CBC + CBC/HB approach (plus one minor difference). First, he collects more information from each choice task (best and worst concepts, plus a more complex None choice). Second, he analyzes the data using purely individual-level estimation rather than HB. As a minor point, he

uses a design methodology that leads to greater level overlap than CBC software's Balanced Overlap approach.

Jordan's main assertion from 2009 was that HB estimation is biased and misleading. To test this claim, I used CBC/HB software to re-analyze the Bottom-Up respondent data for the camera questionnaire. To do so, I coded each choice task as a series of paired comparisons between concepts ("exploded rankings"). The HB run took just 50 minutes for all 600 respondents, even though the rank-order explosion resulted in over 100 choice tasks per respondent. Marshall *et al.* reported the findings, and the HB utilities outperformed the purely individual-level estimation. This clearly demonstrates that (holding the data constant) HB provides better results than the purely individual-level estimation that Louviere implemented in this round of research. In my opinion, Jordan's bold claims in 2009 were wrong. But what about the evidence he reported? One possible explanation is that the version of HB that his team used to analyze the datasets presented in 2009 was faulty, or just didn't use appropriate settings for priors that have proven to work robustly for CBC-type problems.

Is There Value in Best-Worst CBC?

For a few years now, some researchers have advocated asking respondents to identify both the Best and Worst concepts within each choice task (B-W CBC). Three papers at this conference (Chrzan *et al.*, Wirth, and Marshall *et al*.) have presented evidence that asking respondents to identify the worst concept in addition to the best concept can actually improve predictions of best-only holdout choices. Up until this conference, I had been skeptical of the value of asking for worst choices within CBC tasks. This skepticism was a result of research we conducted nearly 15 years ago.

In 1996, Rich Johnson and I re-analyzed a handful of commercial CBC datasets collected under CBC v1 (DOS version). In that first version of CBC, the software permitted the researcher to collect first choice ("best") as well as second choice ("2nd best") third choice ("3rd choice") etc. After the respondent selected the first choice, that concept was removed from the screen, and the respondent was asked to make a selection among the remaining concepts. Rich and I found that data from 2nd choices was not only flatter than first choices (more noisy), but that the utilities were different (after accounting for scale differences). The 2nd choice utilities were biased downward for the best levels of attributes.

After we saw the bias in 2nd choices, we took away the option to ask for anything beyond the best concept in CBC v2 and later.

Marshall *et al.* shared their data for the Bottom-Up respondents, which involved asking respondents to identify both the best and worst concepts in each choice set. I re-analyzed the data for both the pizza and camera datasets using HB estimation. I compared the results of best-only choices to best-worst choices. I found that the inclusion of worst choices damps the scale. But, after tuning the scale up for B-W utilities, I found virtually no difference in the shape of the utility functions when adding worst choices. For these two datasets, there appears to be no bias from adding worst choices to the estimation. Since respondents can supply the additional worst choices very quickly, it might indeed be a good idea.

There are two main differences between the 1996 2nd choice analysis and the one I just conducted on the BU data. First, the 1996 datasets featured minimal overlap. Second, the BU

data involve worst choices rather than 2nd best choices. So, I don't know if the lack of bias for the BU data for worst choices is due to the difference in cognitive process in choosing a worst concept vs. a 2nd best concept, or in the way respondents react to designs featuring minimal overlap vs. moderate overlap. My guess is that the difference lies principally in the focus on worst vs. 2nd best rather than the overlap issue.

We plan to provide an option for asking B-W choices in the next version of our CBC software, so researchers can experiment with this option. Perhaps we'll see more research on this subject in a future Sawtooth Software conference.

At first glance, it doesn't seem logical that adding information regarding worst concepts should help predict what respondents prefer *best* in holdout sets. But, producing a winning concept involves maximizing good aspects and minimizing bad aspects. Thus, considering both kinds of information may be useful in maximizing the likelihood of consumer choice. As long as worst information comes at little cost and is proven to have little or no bias, then it would appear to be a good idea…which gives us another reason to thank Jordan for his contributions. Jordan may not always be right, but he *does* make you think. And, that process can lead to important discovery.

COMMENT ON MARSHALL ET. AL. AND WIRTH

RICHARD T. CARSON, CHRISTINE EBLING,
BART FRISCHKNECHT, JORDAN LOUVIERE AND JOHN ROSE
CENTRE FOR THE STUDY OF CHOICE (CENSOC)

THE CENSOC INDIVIDUAL LEVEL MODELING APPROACH

It seems fair to say that the model-off competition yielded similar performance on most predictive measures for the reigning champion (Hierarchical Bayes, HB) and the CenSoC upstart challenger (individual level models, ILMs). HB and ILMs each claimed some small wins that were offset by wins for the other side. This "draw" should surprise many: HB is a mature technology with a large investment, strong statistical foundation, a large academic literature and wide adoption by marketing practitioners. In contrast, ILMs are essentially new and untested, and fly in the face of decades of conventional wisdom.[1]

We think our first generation ILMs can be improved in several ways, but before discussing this, we outline our basic approach to ensure that it is not misinterpreted. We first estimate parameters for each individual in a sample using individual level weighted least squares.[2] A least squares approach is useful because individual level conditional logit models may not converge for some individuals in a sample. We then predict choice shares by using the estimated WLS regression parameters for each person in a logit link function. The predicted shares and observed 1, 0 choices for each person are used to calculate mean square error between observed and predicted in-sample choice shares. The inverse of mean square error is then used to rescale the original WLS estimates, which has been found to place the subsequent share predictions on a similar scale with the observed shares. We have tested this rescaling approach on many real and simulated data sets, and it gives results similar to those reported at the Conference. We predict out-of-sample aggregate choice shares using the method of sample enumeration, which assumes the sample population represents the population of interest.

For the case of the competition, we asked three questions about each choice set, which gives a partial ordering of the 5 options in each set (A1, A2, A3, A4, and None). Question 1 asks for the most preferred of A1-A4. Question 2 asks for the least preferred of A1-A4. Question 3 asks if the respondent would be satisfied with **any** of the alternatives A1-A4, would choose **some** but not all of the alternatives A1-A4, or would choose **none** of the alternatives. The answer to Question 3 determines the weights used in the WLS regressions; i.e., 5 options have three possible orders, depending on the way one answers the question. The possible weights are 1) Question 3=Any, Ranking: Most=1, Least=3, None=4, Others=2, and associated weights: Most=16, Least=2, None=1, Others=6; 2) Question 3=Some, Ranking: Most=1, Least=3, None=2, Others=2, and associated weights: Most=16, Least=1, None=4.667, Others=4.667; 3) Question 3=None, Ranking: Most=2, Least=4, None=1, Others=3, and associated weights: Most=8, Least=1, None=16, Others=3. We use the weights in WLS regression in the standard way, where the dependent variable is the natural logarithm of the weights (see, e.g., Louviere and

[1] Of course, there is a long history of estimating individual level models in many disciplines including marketing but we appear to be the first to propose their use on a large scale to predict out-of-sample.
[2] An initial sketch of our approach is provided in section 3 of Louviere et al. (2008).

Woodworth 1983). As noted above, we use a logit link function with the estimated parameters to make choice share predictions; we average squared differences between predicted shares and observed 1, 0 choices to get mean squared residual for each person. We reweight each person's WLS parameter estimates using 1/(residual mean square); and use the reweighted parameters with a logit link function to make choice share predictions.

We used the same, fixed design for all respondents; hence, all "see" the same choice sets. Parameter estimates for a particular design may be biased due to omitted effects, but the choice of one design for all respondents avoids design/preference confounds across people associated with random or blocked designs where each person effectively gets a different design or different part of a large design. Our designs try to balance occurrence of attribute levels across sets, not within sets, and to be as efficient as possible according to a D-efficiency criterion (minimize the elements in the model (co)variance matrix).

Our approach can be improved in several ways: 1) Our designs are not optimal for ILMs, and we minimized risk by using designs with more choice sets than theoretically necessary (and more than Sawtooth's HB approach). We now know that we can use designs with the same number of choice sets as Sawtooth, but we have yet to compare their efficiency with the larger designs. We likely can improve our design approach by optimizing other criteria (e.g., prediction accuracy instead of precision). 2) Our ad hoc reweighting seems to work reasonably well in practice, but it is not optimal. 3) We paid little attention to the functional form of the ILMs, which may lead to enhancements. 4) ILMs are "noisy", but there should be ways to cluster people to improve stability and/or use Bayesian methods with "bottom up" estimation.

What Sawtooth Did That Is Germane To Understanding Why They Performed Well

HB (and most choice models) will predict well to any environment with the same error structure. A matched holdout sample is the best case, which is what Wirth simulated, and is effectively what Sawtooth did by "tuning" parameter estimates to in-sample holdout variability. A better standard is to predict real market choices; a less conclusive, but potentially insightful comparison would involve eliciting many more holdout choices.

HB estimation of random coefficient models is one way to get individual-level estimates. HB pulls individual-level estimates toward a sample mean (Allenby & Rossi 1998; Rossi, Allenby & McCulloch 2005). In principle, HB yields individual-level parameter estimates "on the go" and (simulated) maximum likelihood methods estimate individual-level parameters via an extra step that merges information on individual-level observations with aggregate-level model estimates. Huber & Train (2001) show both ways to estimate MIXLs provide similar individual level estimates.

HB typically relies on two priors: a 1st stage prior to define the underlying distribution of heterogeneity (usually assumed normal or lognormal), and a 2nd stage prior ("hyperparameters"). Typically, very diffuse hyperpriors are used to minimize shrinkage in the second stage (Allenby & Rossi 1999). However, as Allenby & Rossi (1999) note: "It is the first stage prior which is important and will always remain important as long as there are only a few observations available" per person. They also note that even if the priors are mis-specified, individual-level posterior means are not constrained to follow this distribution as it is only part of the prior and

individual-level posterior estimates are influenced by the individual estimations. If one has enough observations per person, HB individual estimates should recover the true empirical distribution rather than follow the prior distribution. However, we know of no formal proofs as to what prior distributions are suitable. For example, Allenby & Rossi (1999) only give a diagnostic check for correctness of the underlying distribution assumptions, namely visually checking posterior estimates against prior distributions. We also do not know how many observations per person are "enough".

If HB has "enough" observations and the prior distribution does not impact the individual estimates, one can use the estimates as if they are individual MNLs; and the Swait & Louviere (1993) approach can be used to test scale and preference differences. The latter approach may prove challenging to compare many individuals; so, another way to estimate and compare scales and preferences for individuals is the Louviere & Islam (2008) sequential estimation approach: Use HB to estimate each person's parameters and predict their choice shares, and then calculate their mean squared residuals to estimate their scales. Now use each person's mean-squared residuals as covariates (i.e. interact them with the attribute levels) in a second HB estimation to obtain individual-level preferences adjusted for scale to avoid the scale/preference confound.

RALPH WIRTH'S SIMULATIONS

Ralph did an impressive job, especially considering the scale and scope of his simulation problem. His work highlights several important issues for discrete choice experiments and choice models. We note two issues with his simulations that should be interpreted with caution.

1. The individual model best/worst (BW) methodology

There are several BW models, each with somewhat different properties that can matter in predicting particular types of choices. Ralph focused ONLY on the maximum-difference model, an unrealistic choice process that we find fits less well than others, such as sequential best-worst (Lancsar & Louviere 2008). Thus, simulation results for that model may not generalize to all ILMs or BW models. We have more experience with BW models than ILMs, but there remains much to learn about BW tasks, models and choices, and it is still early days.

2. Changes in scale across samples

Ralph's results suggest HB is unbiased predicting out-of-sample even if individual error variances differ. Communication with Ralph after the Conference suggests it is better to view his findings as providing tentative evidence that HB predicts well when individuals differ in error variances AND in- and out-of-sample error variance distributions are the same. He did not test model performance when error variances differ between samples, a key difference between our ILMs approach and HB. That is, ILMs separate preference parameters and scales, but the algebra of HB clearly shows that HB combines scale and preference differences. Thus, one cannot distinguish people with small (absolute) betas and average scales from people with big (absolute) betas and big scales. Yet, it should be obvious that these two types may have different marketing strategy implications, especially if one can influence noise levels, a common strategy in bundle pricing. Louviere and Eagle (2006) noted this, pointing out that HB will seem to predict well even when model estimates are biased and misleading. Indeed, two lead articles in *Marketing Science* in 2010 (Salisbury & Feinberg; Fiebig, Keane, Louviere & Wasi) give more detail on this issue. Our ILMs approach also faces challenges predicting to cases where noise levels differ

from the estimation data, a situation likely to be the norm, not the exception in marketing applications. However, separating estimates of preference and error variance distributions gives one more options to deal with this challenge.

REFERENCES

Allenby, G. & P. Rossi, P. (1999) Marketing models of consumer heterogeneity. Journal of Econometrics, 89: 57-78.

Fiebig, D., Keane, M., Louviere, J., & Wasi, N. (2010). The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity Marketing Science. 29: 393-421.

Huber, J. & Train, K. (2001). On the similarity of classical and Bayesian estimates of individual mean partworths. Marketing Letters, 12: 259-269.

Lancsar, E. & Louviere, J. (2008). Estimating individual level discrete choice models and welfare measures using best worst choice experiments and sequential best worst MNL, CenSoC Working Paper No. 08-003.

Louviere, J. & Eagle, T. (2006). Confound it! That Pesky Little Scale Constant Messes Up Our Convenient Assumptions. Proceedings of the Sawtooth Software Conference 2006. Sequim, Washington: Sawtooth Software Inc., 211-228.

Louviere, J. & Islam, T. (2008) Separating error variance within & between people: much so-called preference heterogeneity is just due to variability differences in people. Informs Marketing Science Conference 2008, Vancouver.

Louviere, J., Street, D., Burgess, L., Wasi, N., Islam, T. & Marley, A. (2008). Modeling the choices of individual decision-makers by combining efficient choice experiment designs with extra preference information, Journal of Choice Modelling, 1(1): 129-163.

Rossi, P., Allenby, G. & McCulloch, R. (2005) Bayesian statistics and marketing, John Wiley, Hoboken, N.J.

Salisbury, L. & Feinberg, F. (2010). Alleviating the Constant Stochastic Variance Assumption in Decision Research: Theory, Measurement, and Experimental Test, Marketing Science, 29:1-17.

Swait, J. & Louviere, J. (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. Journal of Marketing Research, 30: 305-314.

COMMENT ON MARSHALL ET. AL. AND WIRTH

DAVID W. LYON
AURORA MARKET MODELING, LLC

PROCESS

There is much to admire in the process that culminated in these two papers. Not least is Sawtooth Software's longstanding policy of welcoming (and even soliciting) papers that challenge the ways their software and practitioners, in general, do things. Jordan Louviere's willingness to present just such a paper, boldly claiming superiority for a completely different (and quite new) technique was crucial, of course. Right or wrong, such ideas, forcefully presented with all of Jordan's legendary showmanship, force us to think about topics too often taken for granted. Given Jordan's past contributions to our field, it was only right that his claims be taken seriously and investigated promptly.

Don Marshall, Siu-Shing Chan and Joe Curry responded with a pair of bake-off comparisons, painstakingly designed to be evenhanded. Ralph Wirth performed a truly extensive set of Monte Carlo simulations to systematically explore the effect of a number of dimensions of typical choice problems. Such careful evaluations are particularly remarkable when they come just one conference cycle after Jordan's initial proposal. As Jordan details in his written comments, his ideas have evolved further in the 18 months between conferences. They also did so in the course of the Marshall et al. bake-off (between the pizza study and the later camera study), and will undoubtedly continue to do so in response to the findings of these papers. Nonetheless, a solid read on the performance of the working versions of the idea is far more helpful to us all than postponing evaluations for years in the vain hope that Jordan will run out of more ideas for refinements.

The process culminated at the conference with the presentations of the papers and then with Jordan and Bryan Orme (as spokesperson for the traditional HB MNL approach) confronting and discussing their findings from the same podium. All in all, this was an exemplary case of a new idea being welcomed, proposed, quickly and rigorously tested, and then evaluated and refined.

RESULTS

In the bake-off, the new bottom-up (BU) or individual-level model (ILM) approach did not win. Jordan views the outcome as essentially a draw, but many would call it a loss overall. In Wirth's Monte Carlo work, the traditional HB MNL generally performed better, particularly so in cases of limited data.

But, is this surprising? HB MNL has become the dominant approach over the last 15 or more years, with increasing refinements and understanding of it along the way. The new BU/ILM approach is still being worked on, and the version evaluated in these papers was quite young. In

many ways, the interesting part of the outcome is that a new technique could do so well at such a young age. It too will undergo further refinement, and experience will undoubtedly lead to more effective and efficient implementation and deployment decisions. The story has just begun.

Part of the appeal of BU/ILM is being able to dispense with the complexities of hierarchical Bayes. That idea would have been extremely appealing to practitioners about 15 years ago – we could have avoided learning a new and complex set of techniques. Jordan's deeper point in this regard is that HB incorporates distributional assumptions that he views as directed at the wrong quantities (the utilities, which confound scale and preference) and unjustified in the first place. But from a practical standpoint, most of us have become very comfortable with HB, both in a very practical sense and in terms of the results it produces. Who is to say that utilities are not real, but scale and preference are? Both are only our constructs, after all, not physical constants. And how much should we care about distributional assumptions when elementary expositions of Bayesian ideas usually include demonstrations of just how easily data overwhelms the priors? Claims of cleaner or better theoretical underpinnings relative to HB will not be enough to influence commercial practice on their own.

The downside of BU/ILM is shown in Marshall et al's results on respondent task time and drop-out rates. In addition, keeping modeling self-contained at the individual level imposes minimum data requirements that grow progressively larger as the number of attributes and levels increase. The Marshall et al. camera example was pushing the upper limits of BU/ILM, while many practitioners would call it only a medium-sized problem, or not even quite that. With respondent cooperation continuing its long-term decline, and with practitioners growing ever more adventurous in using very small numbers of choice sets in HB MNL, this downside will be major in commercial practice.

In short, in purely practical terms, there is no compelling reason to shift to BU/ILM modeling today, and some good reasons to resist it. But it is important to keep an eye on future developments and to research further variations. One big area of interest to Jordan is how the psychology of respondent behavior and the mathematics of experimental design interact with each other. We have historically been very focused on the math and to the extent we do consider response psychology, we seldom connect it back to the math. (The work of Keith Chrzan and others on partial profile choice is an obvious, but still very elementary, counterexample.) If Jordan can achieve some breakthroughs here, the result will be better performance, or shorter respondent tasks, or the ability to tackle larger problems, or some of all of those. We can't know now if that will succeed or not, but it is only one potential line for improvements, and Jordan is an inventive and determined visionary leading a large and talented team of researchers.

Revolutions in Market Research – Perspective

Jordan Louviere has already touched off one major revolution in market research, with his tireless proselytizing for choice-based methods in the 1980s. But, it took at least 10 years from his key 1983 paper with Woodworth before we started seeing widespread use of choice models in commercial practice, and perhaps another 5-10 years for them to become clearly dominant over traditional ratings-based conjoint approaches. Greg Allenby and his colleagues had a similarly major impact with hierarchical Bayes methods. But again, 10 years or more elapsed from Greg's

first efforts to tell the HB story to marketing researchers and its ascent to gold standard status in practice.

Watching shooting stars and asteroids is fun. Will they hit or won't they? How big an upheaval will they create? Is this the big one, or just a brief one-time wonder? But in marketing research, unlike in dinosaurs, the asteroids don't cause overnight extinctions. They can cause huge shifts in emphasis, but in slow motion. Has Asteroid Louviere's latest display started a tidal wave we must all either surf or sink into? Perhaps, and it's certainly worth watching, but it will take more than 18 months to tell. In the meantime, I wish it were possible for written proceedings to capture the liveliness and passion of the debate at the conference that these two papers produced and informed.